Statistics and Computing 30 (2020): 1091 – 1111

**DOI:** https://doi.org/10.1007/s11222-020-09937-7

# Optimal Classification of Gaussian Processes in Homo- and Heteroscedastic Settings

**José L. Torrecilla · Carlos Ramos-Carreño · Manuel Sánchez-Montañés · Alberto Suárez**

**Abstract** A procedure to derive optimal discrimination rules is formulated for binary functional classification problems in which the instances available for induction are characterized by random trajectories sampled from different Gaussian processes, depending on the class label. Specifically, these optimal rules are derived as the asymptotic form of the quadratic discriminant for the discretely monitored trajectories in the limit that the set of monitoring points becomes dense in the interval on which the processes are defined. The main goal of this work is to provide a detailed analysis of such optimal rules in the dense monitoring limit, with a particular focus on elucidating the mechanisms by which near perfect classification arises. In the general case, the quadratic discriminant includes terms that are singular in this limit. If such singularities do not cancel out, one obtains near perfect classification, which means that the error approaches zero asymptotically, for infinite sample sizes. This singular limit is a consequence of the orthogonality of the probability measures associated to the stochastic processes from which the trajectories are sampled. As a further novel result of this analysis, we formulate rules to determine whether two Gaussian processes are equivalent or mutually singular (orthogonal).

**Keywords** Functional data analysis - Optimal classification - Gaussian processes - Reproducing kernel Hilbert spaces - Near perfect classification

José L. Torrecilla
[1]Department of Mathematics,
Universidad Autónoma de Madrid, Spain
[2]UC3M-Santander Big Data Institute
E-mail: joseluis.torrecilla@uam.es

Carlos Ramos-Carreño
Computer Science Department,
Universidad Autónoma de Madrid, Spain
E-mail: carlos.ramos@uam.es

Manuel Sánchez-Montañés
Computer Science Department,
Universidad Autónoma de Madrid, Spain
E-mail: manuel.smontanes@uam.es

Alberto Suárez
Computer Science Department,
Universidad Autónoma de Madrid, Spain
E-mail: alberto.suarez@uam.es

## 1 Introduction

Functional data classification is an active research field that has multiple applications in different areas, such as medicine (Zhu et al. 2012; Epifanio and Ventura-Campos 2014), genomics (Leng and Müller 2006; Song et al. 2008; Rincón and Ruiz-Medina 2012), spectrometry (Rossi and Villa 2006), weather modelling and forecasting (Martin-Barragan et al. 2014), speech recognition (Rossi and Villa 2006), and the analysis of handwriting (Hubert et al. 2017). In these types of classification problems, the instances available for induction are characterized by functions of a continuous parameter, such as trajectories in time or curves in space (Ramsay and Silverman 2005; Ferraty and Vieu 2006). Functional classification problems exhibit significant qualitative differences with their multivariate counterparts (Cuevas 2014; Wang et al. 2016). These differences arise from several factors, such as the continuous structure of functions, the lack of a natural ordering in multivariate data, the sparsity in the measurements, and, what is especially relevant for this work, the fact that the densities of random functions do not exist (Delaigle and Hall 2010). In some cases, even if the individual class-conditional probability densities do not exist, an optimal classification rule can still be given in terms of the Radon-Nikodym derivative, which plays

the role of the likelihood ratio in these types of infinite-dimensional problems (Baíllo et al. 2011). In practice, functional data are never complete, in the sense that they are measured only in a grid of points. Furthermore, in some cases the measurements are sparse, which can be a source of additional difficulties (Müller 2016).

In this work we derive explicit expressions of optimal prediction rules for binary classification problems in which the data instances are characterized by trajectories $X$, sampled from Gaussian processes (GPs) defined on the interval $[0, T]$ in the real line. The Gaussian processes from which the instances are sampled are different for each of the two classes. The problem has been analyzed earlier in the literature in both the homo- and heteroscedastic settings (Delaigle and Hall 2012, 2013; Dai et al. 2017; Berrendero et al. 2018b). The main contribution of the current work is to derive such rules by first considering the problem of classifying the discrete-time process that results from monitoring the Gaussian process at a finite set of times. Since the joint distribution of the values of the discretely monitored process is a multivariate Gaussian random variable, the optimal classifier, also called the Bayes rule, is a quadratic discriminant (Hastie et al. 2009). By taking the limit of this quadratic discriminant as the set of monitoring points becomes dense in $[0, T]$, one obtains an optimal rule for the classification of the continuous-time Gaussian processes. In the general case, this is a singular limit because some of the terms in the discriminant rule diverge. Carrying out a detailed analysis of these optimal classification rules and their singularities in the dense monitoring limit, we provide novel derivations of some known results and gain insight into the mechanisms by which near perfect classification occurs (Delaigle and Hall 2012). Specifically, for these types of problems, an optimal classification rule is obtained by balancing the singular terms that appear in the quadratic discriminant in the dense monitoring limit. The conditions for near perfect classification when the Gaussian processes have different means were first derived in Delaigle and Hall (2012) in the homo- and heteroscedastic settings. The current paper builds on that work by considering the covariance structure as well. A further novel result is the derivation of explicit rules to determine whether two Gaussian processes are equivalent or mutually singular (Hájek 1958; Feldman 1958). Two processes are said to be equivalent when their zero measure sets coincide. They are mutually singular (orthogonal) when there is a non-empty set that has zero measure under one process and measure one for the other one. The equivalence rules are derived from the observation that for the two processes to be equivalent, the singularities that appear in the corresponding classification rule must cancel out.

The structure of the article is as follows: The functional classification problem is formulated in Section 2. A natural framework for the analysis of these types of problems is the theory of Reproducing Kernel Hilbert Spaces (RKHS's), which is reviewed in Section 3. The procedure for deriving optimal rules for the classification of Gaussian processes based on discrete monitoring is introduced in Section 4. The type of classification problem that is obtained depends on whether the Gaussian processes are equivalent (non-singular classification) or orthogonal (singular, near perfect classification). For this reason, the conditions for the equivalence of Gaussian processes are discussed in Section 5. Sections 6 and 7 are devoted to homoscedastic and heteroscedastic classification problems, respectively. An experimental evaluation of the limit rules derived in this work and a comparison with other functional classification methods is presented in Section 8 for both simulated and real-world problems. Finally, Section 9 provides a summary of the conclusions of this work.

## 2 Statement of the problem

In functional classification, the instances available for learning are characterized by pairs $(X, Y)$, where $X$ is a function of the continuous parameter $t \in \mathcal{I}$ and $Y$ is a discrete class label. In a binary classification problem $Y \in \{0, 1\}$. For the sake of simplicity, we assume that $\mathcal{I}$ is a compact interval in the real line (e.g., $t \in [0, T]$). Nonetheless, the results can be readily extended to cases in which $\mathcal{I}$ is a compact domain in a Euclidean space of arbitrary dimension. Assuming that we have at our disposal a set of labeled training examples $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N_{train}}$, the goal is to induce from these data a predictor that, given a known $x$ as input, produces a class label as output.

Assume that the functions $x \equiv \{x(t); t \in \mathcal{I}\}$ are realizations of the stochastic process

$$\begin{aligned} (X(t) \mid Y = 0) &= m_0(t) + Z_0(t) \qquad \text{w. p. } 1 - p \\ (X(t) \mid Y = 1) &= m_1(t) + Z_1(t) \qquad \text{w. p. } p, \end{aligned} \tag{1}$$

where $p$ is the prior of class 1 ($0 < p < 1$), and $m_0(t)$, $m_1(t)$ are the mean of class 0 and class 1 instances, respectively. Both means are square-integrable deterministic functions on $\mathcal{I}$

$$\int_{t \in \mathcal{I}} |m_i(t)|^2 \, dt < \infty, \quad i = 0, 1. \tag{2}$$

The stochastic terms, $Z_0$ and $Z_1$, are assumed to be zero-mean Gaussian processes with continuous covariance functions (kernels) $K_0$ and $K_1$, respectively.

Without loss of generality, the mean $m_0$ can be subtracted from all trajectories to obtain the equivalent classification problem

$$\begin{aligned}(X(t) \mid Y=0) &= Z_0(t) && \text{w. p. } 1-p \\ (X(t) \mid Y=1) &= m(t)+Z_1(t) && \text{w. p. } p,\end{aligned} \tag{3}$$

where $m(t) = m_1(t) - m_0(t)$.

The measures $\mathbb{P}_0$ and $\mathbb{P}_1$ are the laws of the stochastic processes $X_0 = (X \mid Y = 0)$ and $X_1 = (X \mid Y = 1)$, respectively. For Gaussian processes, Hájek (1958) and Feldman (1958) established that these measures are either equivalent or mutually singular. Necessary and sufficient conditions for the equivalence of Gaussian measures have been given in the literature (Varberg 1961; Parzen 1961b,a; Shepp 1966; Sato 1967; Kuelbs 1970).

In the case that $\mathbb{P}_0$ and $\mathbb{P}_1$ are equivalent ($\mathbb{P}_0 \sim \mathbb{P}_1$), the Radon-Nikodym derivative between the two measures $\frac{d\mathbb{P}_1}{d\mathbb{P}_0}(X)$, which is the analogue of the likelihood ratio for infinite-dimensional functional spaces, exists. Furthermore, the Bayes rule (i.e., the optimal classification rule) can be expressed in terms of this derivative (Baíllo et al. 2011). Specifically, the optimal predictor for an instance characterized by the trajectory $x$ is

$$\mathbb{I}\left[\left.\frac{d\mathbb{P}_1}{d\mathbb{P}_0}(X)\right|_{X=x} > \frac{1-p}{p}\right], \tag{4}$$

where $\mathbb{I}$ is the indicator function (i.e., $\mathbb{I}[True] = 1$, $\mathbb{I}[False] = 0$). When the two measures are mutually singular ($\mathbb{P}_0 \perp \mathbb{P}_1$), near perfect classification is obtained (Berrendero et al. 2018b). Near perfect classification was first discussed in (Delaigle and Hall 2012). In that paper the authors showed that zero classification error can be achieved in the asymptotic (infinite sample) limit for Gaussian processes with different means that fulfill certain conditions both in the homo- and heteroscedastic settings. In the current article, we build on this seminal work and provide detailed derivations of optimal classification rules that illustrate the emergence of the near perfect classification phenomenon in different cases, including those involving the covariances of the Gaussian processes.

The expression of the Bayes rule given by Eq. (4) is formal and cannot be directly used in practical applications. Approximations based on the use of density ratios of finite dimensional projections have been proposed in Delaigle and Hall (2013); Galeano et al. (2015); Dai et al. (2017). The derivation of explicit forms for the optimal rule for a limited class of functional classification problems of this type has also been considered earlier in the literature mainly in the homoscedastic ($K_0 = K_1 = K$) setting. However, the derivation of optimal classification rules in the heteroscedastic setting ($K_0 \neq K_1$) and for singular cases in which near perfect classification is obtained remains elusive (Delaigle and Hall 2012, 2013; Cuesta-Albertos and Dutta 2016; Dai et al. 2017; Berrendero et al. 2018b). In a homoscedastic setting, some related results, including the singular case, have been derived in the context of signal detection (Kailath 1966, 1971).

A natural framework for these types of functional classification problems is the theory of Reproducing Kernel Hilbert Spaces (Cucker and Smale 2002; Berlinet and Thomas-Agnan 2004; Manton and Amblard 2015). For this reason, in the next section we provide an overview of properties of these types of spaces that will be used later in this article to derive optimal rules for the classification of Gaussian processes.

## 3 Reproducing Kernel Hilbert Spaces

A Reproducing Kernel Hilbert Space (RKHS) is a space of real-valued functions on $\mathcal{I}$ endowed with an inner product that is a reproducing kernel. A kernel $K(s,t)$ is a symmetric positive-semidefinite function on $\mathcal{I} \times \mathcal{I}$. The reproducing property for the kernel $K$ associated with the RKHS $\mathcal{H}_K$ implies that

$$f(t) = \langle f(\cdot), K(\cdot,t)\rangle_K \ , \forall f \in \mathcal{H}_K, \tag{5}$$

where $\langle\cdot,\cdot\rangle_K$ denotes the inner product in $\mathcal{H}_K$.

Throughout this work we will assume that $K$ is continuous and bounded. The positivity condition implies that for any finite set of distinct values $\{t_n\}_{i=1}^N \in \mathcal{I}^N$ the $N \times N$ Gram matrix $\mathbf{K}$, whose elements are $K_{nm} = K(t_n, t_m)$, for $1 \leq n, m \leq N$, is positive-semidefinite (i.e. all its eigenvalues are non-negative).

Let $\mathcal{H}_K^{(0)}$ be the space of functions that can be expressed as finite linear combinations of the form $f(\cdot) = \sum_{n=1}^r \alpha_n K(t_n, \cdot)$. This space is endowed with the inner product

$$\langle f, g\rangle_K = \sum_{n=1}^{r}\sum_{m=1}^{s} \alpha_n \beta_m K(t_n, t_m), \tag{6}$$

where $g(\cdot) = \sum_{m=1}^s \beta_m K(t_m, \cdot)$. The RKHS associated to $K$, denoted by $\mathcal{H}_K$, is the set of functions $f : \mathcal{I} \to \mathbb{R}$ that is the completion in the corresponding norm of the space $\mathcal{H}_K^{(0)}$.

Let $L^2(\mathcal{I})$ denote the space of square integrable functions on $\mathcal{I}$. Associated to kernel $K$, it is possible to define a covariance operator $\mathcal{K} : L^2(\mathcal{I}) \to L^2(\mathcal{I})$, by the integral equation

$$\mathcal{K}f(t) = \int_{s\in\mathcal{I}} K(t,s)f(s)ds. \tag{7}$$

This operator is self-adjoint and positive. The eigenvalues and eigenfunctions of this operator are

$$\int_{s\in\mathcal{I}} K(t,s)\phi_j(s)ds = \lambda_j\phi_j(t), \quad j=1,2,\dots \tag{8}$$

with $0 < \dots \leq \lambda_2 \leq \lambda_1 < \infty$, and $\{\phi_j(s)\}_{j=1}^{\infty}$, an orthonormal basis in $L^2(\mathcal{I})$

$$\int_{t\in\mathcal{I}} \phi_i(t)\phi_j(t)dt = \delta_{ij}, \quad i,j=1,2,\dots. \tag{9}$$

If the kernel is only positive-semidefinite, some of the eigenvalues of $\mathcal{K}$ could be zero. In that case, there is no loss of generality in considering only the linear subspace spanned by the set of eigenvectors corresponding to positive (non-zero) eigenvalues of the covariance operator (see, e.g., *Remark 3* of Section 3 in Cucker and Smale (2002)).

By Mercer's theorem (Parzen 1959; Cucker and Smale 2002), the spectral representation of the kernel is

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i\phi_i(s)\phi_i(t). \tag{10}$$

The convergence of this series is absolute for each $(s,t) \in \mathcal{I}\times\mathcal{I}$ and uniform on $\mathcal{I}\times\mathcal{I}$. In this representation, the RKHS associated to kernel $K$ is the space of functions that fulfill

$$\mathcal{H}_K = \left\{ f \in L^2(\mathcal{I}) :\ f = \sum_{i=1}^{\infty} f_i\phi_i,\ \sum_{i=1}^{\infty} \frac{f_i^2}{\lambda_i} < \infty \right\}. \tag{11}$$

The corresponding inner product is

$$\langle f,g\rangle_K = \sum_{i=1}^{\infty} \frac{f_i g_i}{\lambda_i}, \tag{12}$$

with $g = \sum_{i=1}^{\infty} g_i\phi_i(t)$. Finally, the set of functions $\left\{\varphi_j(s) = \sqrt{\lambda_j}\phi_j(s)\right\}_{j=1}^{\infty}$ form an orthonormal basis in $\mathcal{H}_K$.

### 3.1 Stochastic processes and RKHS's

There is an alternative construction of the RKHS that takes as a starting point the zero-mean second-order stochastic process $\{Z(t), t\in\mathcal{I}\}$, whose covariance function is

$$K(s,t) = \mathbb{E}\left[Z(s)Z(t)\right], \quad s,t\in\mathcal{I}. \tag{13}$$

Consider $\mathcal{L}_0(Z)$, the linear span of the process $Z$, whose elements are of the form $\sum_{n=1}^{N}\alpha_n Z(t_n)$, with $\{\alpha_n\}_{n=1}^{N} \in \mathbb{R}^N$, $\{t_n\}_{n=1}^{N} \in \mathcal{I}^N$ for some integer $N$. Let $\mathcal{L}(Z)$ be the closure of $\mathcal{L}_0(Z)$ in $L^2(\Omega)$, the space of zero-mean random variables with finite second moments. By Loève's representation theorem (Berlinet and Thomas-Agnan 2004) it is possible to define an isomorphism that maps each $\sum_n \alpha_n Z(t_n) \in \mathcal{L}(Z)$ onto a unique element

$$\begin{aligned}\psi\left(\sum_n \alpha_n Z(t_n)\right)(t) &= \mathbb{E}\left[\sum_n \alpha_n Z(t_n)Z(t)\right] \\ &= \sum_n \alpha_n K(t_n,t), \quad t\in\mathcal{I}\end{aligned} \tag{14}$$

in $\mathcal{H}_K$. Since $\sum_n \alpha_n K(t_n,t)$ converges in the norm sense to an element of the RKHS (because $\sum_n \alpha_n Z_n$ belongs to the closure of $\mathcal{L}_0(Z)$ in $L^2(\Omega)$), it also converges pointwise for all $t\in\mathcal{I}$ to the same limit (Corollary 1 of Berlinet and Thomas-Agnan (2004)). This isomorphism preserves the inner product; i.e., it is a congruence. This congruence is referred to as Loève's isometry (Lukić and Beder 2001). It maps $Z(s)$ onto $K(s,t)$

$$\psi(Z(s))(t) = K(s,t), \quad s,t\in\mathcal{I}. \tag{15}$$

Conversely, the inverse congruence $\psi_Z^{-1}$ maps the function $f = \sum_n \alpha_n K(t_n,\cdot) \in \mathcal{H}_K$ onto the random variable $\psi_Z^{-1}(f) = \sum_n \alpha_n Z(t_n) \in \mathcal{L}(Z)$. Therefore, the value of the random trajectory at $t\in\mathcal{I}$ can be expressed as

$$Z(t) = \psi_Z^{-1}\left(K(\cdot,t)\right). \tag{16}$$

In terms of this isometry, the inner product between the functions $f = \sum_n \alpha_n K(t_n,\cdot)$ and $g = \sum_m \beta_m K(t_m,\cdot)$, both in $\mathcal{H}_K$ is

$$\langle f,g\rangle_K = \sum_{n,m} \alpha_n\beta_m K(t_n,t_m) = \mathbb{E}\left[\psi_Z^{-1}(f)\psi_Z^{-1}(g)\right]. \tag{17}$$

Following Parzen (1961a), this isometry can be used to define the mapping

$$\langle Z,f\rangle_K = \psi_Z^{-1}(f) = \sum_n \alpha_n Z(t_n) \in \mathcal{L}(Z), \tag{18}$$

for $Z$, a trajectory of the stochastic process, and $f\in\mathcal{H}_K$. This mapping, which can be viewed as a formal extension of the definition of the inner product, is well-defined even though, except for trivial cases, $Z \notin \mathcal{H}_K$ with probability one (Kailath 1971; Berlinet and Thomas-Agnan 2004). The quantity $\langle Z,f\rangle_K$ is the unique linear square-integrable functional of $Z$ that satisfies (Parzen 1959; Kailath 1971)

$$\mathbb{E}\left[Z\langle Z,f\rangle_K\right] = f,\ \forall f\in\mathcal{H}_K. \tag{19}$$

Finally, it is also possible to express this *congruence* inner product as

$$\langle Z,f\rangle_K = \sum_{i=1}^{\infty} \frac{\zeta_i f_i}{\lambda_i}, \tag{20}$$

in terms of $f(t) = \sum_{i=1}^{\infty} f_i \phi_i(t),\ t \in \mathcal{I}$, and of $Z(t) = \sum_{i=1}^{\infty} \zeta_i \phi_i(t),\ t \in \mathcal{I}$, the Karhunen-Loève expansion of the process (Berlinet and Thomas-Agnan 2004). The coordinates of this expansion are computed by projecting $Z$ onto the basis of eigenfunctions of $\mathcal{K}$

$$\zeta_i = \int_{t \in \mathcal{I}} Z(t)\phi_i(t)dt,\ i = 1, 2, \ldots. \tag{21}$$

They are zero-mean independent normal random variables

$$\begin{aligned} &\mathbb{E}\left[\zeta_i\right] = 0 \\ &\mathbb{E}\left[\zeta_i \zeta_j\right] = \lambda_i \delta_{ij}, \quad i, j = 1, 2, \ldots. \end{aligned} \tag{22}$$

In the following section, these properties will be used to derive rules for the optimal classification of trajectories sampled from two different Gaussian processes.

## 4 Optimal rules for Gaussian process classification

Consider a stochastic process $X$ defined by Eq. (3). Assume that the trajectories of this process are monitored at a set of appropriately chosen distinct discrete times $\mathbf{t}_N = \{t_n\}_{n=1}^N \in \mathcal{I}^N$. Let $\mathbf{X}$ represent the $N$-dimensional column vector whose components are the discretely monitored values of the trajectories

$$\mathbf{X}^\dagger = \left(X(t_1),\, X(t_2),\, \ldots,\, X(t_N)\right), \tag{23}$$

where the superscript † indicates the standard transposition of matrices. By the properties of Gaussian processes, the class-conditioned distribution of $\mathbf{X}$ is a multivariate Gaussian

$$\begin{aligned} &\mathbf{X} \mid Y = 0 \sim N\left(\mathbf{0}, \mathbf{K}_0\right) && \text{w. p. } 1 - p \\ &\mathbf{X} \mid Y = 1 \sim N\left(\mathbf{m}, \mathbf{K}_1\right) && \text{w. p. } p \end{aligned} \tag{24}$$

where

$$\mathbf{m}^\dagger = \left(m(t_1),\, m(t_2),\, \ldots,\, m(t_N)\right) \tag{25}$$

is a row vector whose components are the values of mean of the class 1 trajectories at the monitoring times. The corresponding column vector is denoted by $\mathbf{m}$.

The quantities $\mathbf{K}_0$ and $\mathbf{K}_1$ are the corresponding $N \times N$ Gram matrices. The elements of these matrices are the autocovariances of the discretely monitored processes

$$\left(\mathbf{K}_i\right)_{mn} = \mathbb{E}\left[Z_i(t_n) Z_i(t_m)\right] = K_i(t_n, t_m), \tag{26}$$

for $i = 0, 1$, and $n, m = 1, 2, \ldots N$. Since they characterize the structure of autocovariances, Gram matrices are positive-semidefinite (i.e., their eigenvalues are non-negative). If they have zero eigenvalues, in what follows, the derivations apply to the space spanned by the eigenvectors corresponding to the positive (non-zero) eigenvalues.

In the general heteroscedastic case, the Bayes rule for this multivariate Gaussian binary classification problem is the quadratic discriminant (see, e.g., Hastie et al. (2009))

$$\begin{aligned} \mathbb{I}\Bigg[ &-\frac{1}{2}\log\frac{|\mathbf{K}_1|}{|\mathbf{K}_0|} - \frac{1}{2}\mathbf{X}^\dagger\left(\mathbf{K}_1^{-1} - \mathbf{K}_0^{-1}\right)\mathbf{X} \\ &+ \mathbf{X}^\dagger \mathbf{K}_1^{-1}\mathbf{m} - \frac{1}{2}\mathbf{m}^\dagger \mathbf{K}_1^{-1}\mathbf{m} > \log\frac{1-p}{p}\Bigg], \end{aligned} \tag{27}$$

where $\mathbb{I}$ is the indicator function and $|\mathbf{K}_0|, |\mathbf{K}_1|$ are the determinants of the corresponding covariance matrices.

The limit of this rule as $N \to \infty$ and the set of monitoring points $\mathbf{t}_\infty = \{t_n\}_{n=1}^\infty$ becomes dense in $\mathcal{I}$ can be formally written as

$$\begin{aligned} \mathbb{I}\Bigg[ &-\frac{1}{2}\log\frac{|\mathcal{K}_1|}{|\mathcal{K}_0|} - \frac{1}{2}\left(\langle X, X\rangle_{K_1} - \langle X, X\rangle_{K_0}\right) \\ &+ \langle X, m\rangle_{K_1} - \frac{1}{2}\langle m, m\rangle_{K_1} > \log\frac{1-p}{p}\Bigg]. \end{aligned} \tag{28}$$

The angular brackets $\langle\cdot,\cdot\rangle_{K_i}$ denote the inner product in $\mathcal{H}_i$, the reproducing kernel Hilbert spaces (RKHS) associated to the kernel $K_i$, or, if such quantity is ill-defined, related mathematical constructs whose particular form will be made explicit later on in this work. The quantity $\frac{|\mathcal{K}_1|}{|\mathcal{K}_0|}$ represents the asymptotic form of the ratio of determinants of the Gram matrices $\frac{|\mathbf{K}_1|}{|\mathbf{K}_0|}$ in the limit of dense monitoring (see Appendix A).

In general cases, this limit is singular. A first type of singularity occurs if $m \notin \mathcal{H}_0 \cap \mathcal{H}_1$. In such case, the terms $\langle X, m\rangle_{K_1}$ and $\langle m, m\rangle_{K_1}$ in Eq. (28) diverge. A second type of singularity appears in the quadratic terms of the discriminant when the Hilbert space $\mathcal{H}_i$ is infinite-dimensional. In that case, the trajectories of the process $X$ do not belong to $\mathcal{H}_K$ with probability one (Kailath 1971; Berlinet and Thomas-Agnan 2004). Therefore, in the dense monitoring limit, the quantities $\langle X, X\rangle_{K_i}$ also diverge. Finally, also for infinite-dimensional $\mathcal{H}_i$, the determinant of the corresponding covariance operator, $|\mathcal{K}_i|$, vanishes and its logarithm diverges.

As illustrated in this work, these singularities are in fact at the origin of the near perfect classification phenomenon (Delaigle and Hall 2012). Specifically, if the singularities present in the classification rule Eq. (28) cancel out, the measures of the two underlying Gaussian processes are equivalent. Otherwise, they are mutually singular (orthogonal) and near perfect classification is obtained.

From these observations, we note that Eq. (28), which can be seen as the functional generalization of the quadratic discriminant for multivariate data, should be viewed only as a mnemonic for Eq. (27) in the limit of

dense monitoring. In subsequent sections, the singular limit of this rule is analyzed in detail for different classification problems in both the homo- and heteroscedastic settings. In particular, the conditions for the equivalence between the two Gaussian processes are discussed in Section 5. In section 6 we analyze homoscedastic classification problems, for which $K_0 = K_1 = K$. In this case, if $m = m_1 - m_0 \in \mathcal{H}_K$, the laws $\mathbb{P}_0$ and $\mathbb{P}_1$ are equivalent. In consequence, the classification problem is not singular and Eq. (28) is the Bayes rule (Berrendero et al. 2018b). Near perfect classification is obtained when $m$ does not belong to $\mathcal{H}_K$. In such case the singularities in the terms that involve $m$ dominate in Eq. (28). Nonetheless, it is still possible to derive optimal classification rules by carrying out a careful analysis of the behavior of those divergent terms in this singular limit. The general heteroscedastic classification problem $K_0 \neq K_1$ is analyzed in Section 7. Near perfect classification can be obtained by an alternative mechanism that involves the quadratic terms of the discriminant, which are singular. As in the homoscedastic case, if the singularities in Eq. (28) cancel out, $\mathbb{P}_0$ and $\mathbb{P}_1$ are equivalent, and the classification problem is not singular.

## 5 Equivalence of Gaussian processes

From the form of the optimal classification rule introduced in the previous section it is possible to derive conditions for the equivalence of the probability measures $\mathbb{P}_0$ and $\mathbb{P}_1$ associated to the Gaussian processes $GP(m_0, K_0)$ and $GP(m_1, K_1)$, respectively.

The derivation starts from the observation that $\mathbb{P}_0$ and $\mathbb{P}_1$ are equivalent if the corresponding classification problem is not singular (Baíllo et al. 2011; Berrendero et al. 2018b). In that case, the optimal classification rule is

$$\mathbb{I}\left[\frac{d\mathbb{P}_1}{d\mathbb{P}_0}(X) > \frac{1-p}{p}\right], \tag{29}$$

where $\frac{d\mathbb{P}_1}{d\mathbb{P}_0}(X)$ is the Radon-Nikodym derivative.

If $m_0 \neq 0$ it is possible to determine whether the trajectory $\{X(t); t \in \mathcal{I}\}$ has been sampled either from $GP(m_0, K_0)$ or from $GP(m_1, K_1)$ using Eq. (28) with the replacements $X - m_0$ for $X$ and $m = m_1 - m_0$. The classification rule becomes

$$\mathbb{I}\Bigg[-\frac{1}{2}\log\frac{|\mathcal{K}_1|}{|\mathcal{K}_0|} - \frac{1}{2}\left(\|X-m_1\|^2_{K_1} - \|X-m_0\|^2_{K_0}\right) > \log\frac{1-p}{p}\Bigg], \tag{30}$$

where $\|X - m_i\|^2_{K_i} = \langle X - m_i, X - m_i\rangle_{K_i}$, with $i = 0, 1$, which are the functional analogues of the Mahalanobis distance (Galeano et al. 2015; Berrendero et al. 2018a). Again, the quantities that appear in this expression exhibit singularities and should therefore be interpreted as the corresponding discrete approximations in the limit of dense monitoring.

A first type of singularity in this classification rule arises when $m = m_1 - m_0 \notin \mathcal{H}_0 \cap \mathcal{H}_1$. Specifically, assuming that $X$ is a trajectory drawn from $GP(m_1, K_1)$, it can be written as $X = m_1 + Z_1$, where $Z_1$ is the zero-mean Gaussian process $GP(0, K_1)$. In the expression of $\|X - m_0\|^2_{K_0}$ one would get, among others, the term $\|m_1 - m_0\|^2_{K_0}$, which diverges because $m = m_1 - m_0 \notin \mathcal{H}_0$. Note that this singularity cannot be cancelled by any other term in Eq. (30). A parallel argument can be used for trajectories drawn from $GP(m_0, K_0)$, interchanging the subindices 0 and 1, to prove that, if $m = m_1 - m_0 \notin \mathcal{H}_1$, one would get the term $\|m_1 - m_0\|^2_{K_1}$ in Eq. (30). This term is also singular because $m = m_1 - m_0 \notin \mathcal{H}_1$. As in the previous case, the singularity cannot be cancelled by any other term in Eq. (30).

Even if one assumes that $m = m_1 - m_0 \in \mathcal{H}_0 \cap \mathcal{H}_1$, which implies that the divergences described in the previous paragraph are not obtained, a second type of singularity can appear in the quadratic terms of Eq. (30). Specifically, the terms $\|X - m_i\|^2_{K_i}$, $i = 0, 1$ diverge when $\mathcal{H}_i$ is infinite dimensional (Berrendero et al. 2018a). However, if the Gaussian processes are equivalent, the singularities in the term $\left(\|X - m_1\|^2_{K_1} - \|X - m_0\|^2_{K_0}\right)$ cancel out, so that the classification rule given by Eq. (30) is well defined.

A related singularity affects also the term that involves the ratio of the determinants of the covariance operators. If the Hilbert space $\mathcal{H}_i$ is infinite dimensional, zero is an accumulation point of the spectrum of the covariance operator $\mathcal{K}_i$ (Spence 1975). Therefore, the individual determinants of the covariance operators vanish

$$|\mathcal{K}_i| \equiv \lim_{N\to\infty}\prod_{j=1}^{N}\lambda_{ij} = 0, \quad i = 0, 1, \tag{31}$$

where $\{\lambda_{0j}\}_{j=1}^{\infty}$ and $\{\lambda_{1j}\}_{j=1}^{\infty}$ are the eigenvalues of $\mathcal{K}_0$ and $\mathcal{K}_1$, respectively. Therefore, the condition for equivalence between $\mathbb{P}_0$ and $\mathbb{P}_1$ is that the ratio determinants

$$\frac{|\mathcal{K}_0|}{|\mathcal{K}_1|} = \lim_{N\to\infty}\frac{|\mathbf{K}_0|}{|\mathbf{K}_1|} = \lim_{N\to\infty}\prod_{j=1}^{N}\frac{\lambda_{0j}}{\lambda_{1j}}, \tag{32}$$

be finite and different from zero. The last equality in Eq. (32), which involves the limit of dense monitoring, is derived in Appendix A.

In summary, if the processes are equivalent, the classification problem is not singular. Therefore,

$$m = m_1 - m_0 \in \mathcal{H}_0 \cap \mathcal{H}_1, \tag{33}$$

so that the terms that depend solely on the means in Eq. (30) are well-defined, and the singularities of the quadratic terms cancel out

$$\|X - m_1\|^2_{K_1} - \|X - m_0\|^2_{K_0} < \infty, \tag{34}$$

$$0 < \frac{|\mathcal{K}_0|}{|\mathcal{K}_1|} = \lim_{N\to\infty} \prod_{j=1}^{N} \frac{\lambda_{0j}}{\lambda_{1j}} < \infty. \tag{35}$$

If these conditions do not hold, the classification problem is singular and the measures are not equivalent. In consequence, according to the Hájek-Feldman dichotomy (Hájek 1958; Feldman 1958), they are mutually singular (orthogonal).

For processes that are equivalent, combining Eqs. (29) and (30), the Radon-Nikodym derivative of $\mathbb{P}_1$ with respect to $\mathbb{P}_0$ is

$$\frac{d\mathbb{P}_1}{d\mathbb{P}_0}(x) = \left(\frac{|\mathcal{K}_0|}{|\mathcal{K}_1|}\right)^{1/2} \exp\left\{-\frac{1}{2}\left(\|x - m_1\|^2_{K_1} - \|x - m_0\|^2_{K_0}\right)\right\}. \tag{36}$$

Furthermore, the inner products in $\mathcal{H}_0$ and $\mathcal{H}_1$, the RKHS's corresponding to the kernels $K_0$ and $K_1$, respectively, are related by the expression

$$\langle f, g\rangle_{K_1} = \langle f, g\rangle_{K_0} - \langle f, \langle \delta K, g\rangle_{K_1}\rangle_{K_0}, \quad f, g \in \mathcal{H}_0 \cap \mathcal{H}_1, \tag{37}$$

where $\delta K = K_1 - K_0$. This relation can be proven making use of the reproducing property of $K_0$ in $\mathcal{H}_0$, and of $K_1$ in $\mathcal{H}_1$:

Let $f \in \mathcal{H}_0 \cap \mathcal{H}_1$. Using $g(\cdot) = K_1(x, \cdot)$ in Eq. (37)

$$\begin{aligned}
&\langle f(\cdot), K_1(x,\cdot)\rangle_{K_1} \\
&\quad = \langle f(\cdot), K_1(x,\cdot)\rangle_{K_0} - \langle f(\cdot), \langle \delta K(\cdot,\cdot), K_1(x,\cdot)\rangle_{K_1}\rangle_{K_0} \\
&\quad = \langle f(\cdot), K_1(x,\cdot)\rangle_{K_0} - \langle f(\cdot), \delta K(x,\cdot)\rangle_{K_0} \\
&\quad = \langle f(\cdot), K_0(x,\cdot)\rangle_{K_0} = f(x).
\end{aligned}$$

Using these results for equivalence, we now proceed to analyze the classification of Gaussian processes in both the homo- and heteroscedastic settings.

## 6 Homoscedastic classification problems

In homoscedastic classification problems, the kernels of the Gaussian processes for the two classes are equal $K_0 = K_1 = K$. Therefore, the quadratic terms of the functional discriminant function (i.e., the first two terms on the left-hand side of the expression inside the indicator function in Eq. (27)) cancel out. The optimal rule for the discretely monitored process is Fisher's linear discriminant

$$\mathbb{I}\left[\mathbf{X}^\dagger \mathbf{K}^{-1}\mathbf{m} - \frac{1}{2}\mathbf{m}^\dagger \mathbf{K}^{-1}\mathbf{m} > \log\frac{1-p}{p}\right]. \tag{38}$$

In the limit of dense monitoring the decision rule can be formally written as

$$\mathbb{I}\left[\langle X, m\rangle_K - \frac{1}{2}\langle m, m\rangle_K > \log\frac{1-p}{p}\right]. \tag{39}$$

The angular brackets, $\langle\cdot,\cdot\rangle_K$ denote the inner product in $\mathcal{H}_K$, the RKHS associated to the kernel $K$, or, if such quantity is ill-defined, a related mathematical construct that will be described presently.

In what follows, the Bayes rule will be derived for general non-singular homoscedastic GP classification problems. Then we will illustrate how to derive optimal rules in specific singular instances of such problems.

### 6.1 Non-singular homoscedastic classification

In a homoscedastic setting, the classification problem is not singular if $m = m_1 - m_0 \in \mathcal{H}_K$ (lemma 5d of Parzen (1961b)). In that case, $\mathbb{P}_0$ and $\mathbb{P}_1$ are equivalent. Provided that an appropriate interpretation is given to its constituents, Eq. (39) is the Bayes rule for this functional classification problem. The error of this optimal rule, which is the infinite-dimensional analogue of Fisher's linear discriminant, is

$$\begin{aligned}
L^* = (1-p)\Phi\left(-\tfrac{1}{2}\|m\|_K - \tfrac{1}{\|m\|_K}\log\tfrac{p}{1-p}\right) \\
+ p\Phi\left(-\tfrac{1}{2}\|m\|_K + \tfrac{1}{\|m\|_K}\log\tfrac{p}{1-p}\right),
\end{aligned} \tag{40}$$

where $\Phi$ is the cumulative distribution function of a standard normal random variable (Berrendero et al. 2018b).

As mentioned earlier, the terms in Eq. (39) need to be given an appropriate interpretation in the limit of dense monitoring. Let's consider first the term that involves the inner product of $m$

$$\lim_{N\to\infty} \mathbf{m}^\dagger \mathbf{K}^{-1}\mathbf{m} = \langle m, m\rangle_K = \|m\|^2_K. \tag{41}$$

The convergence of the discretized approximation to the square norm of $m \in \mathcal{H}_K$ can be proven for monotone increasing (nested) sets of monitoring times using lemma 5c of Parzen (1961b).

If the RKHS is infinite dimensional, the trajectories of the random process do not belong to $\mathcal{H}_K$ with probability one (Kailath 1971; Lukić and Beder 2001; Berlinet and Thomas-Agnan 2004). In such case, $\langle X, m\rangle_K$ cannot

represent an inner product in $\mathcal{H}_K$. Nonetheless, since $m \in \mathcal{H}_K$, we have

$$\mathbb{E}\left[Z(t_n)\,\left(\mathbf{X}^\dagger\mathbf{K}^{-1}\mathbf{m}\right)\right] = m(t_n),\ t_n \in \mathbf{t}_N \tag{42}$$

$$\mathbb{E}\left[Z(t)\,\left(\mathbf{X}^\dagger\mathbf{K}^{-1}\mathbf{m}\right)\right] = \hat{m}(t),\ t \notin \mathbf{t}_N, \tag{43}$$

where $Z(t)$ is a random function sampled from a zero-mean Gaussian process with a kernel function $K$. The quantity $\hat{m}(t)$ is the optimal prediction for $m(t)$ with $t \in \mathcal{I}$, assuming that $\{m(t_n)\}_{n=0}^N$, the values of the mean at the monitoring times, are known (Rasmussen and Williams 2005). In the limit of dense monitoring

$$\lim_{N\to\infty} \mathbb{E}\left[Z(t)\left(\mathbf{X}^\dagger\mathbf{K}^{-1}\mathbf{m}\right)\right] = m(t), \quad \forall t \in \mathbf{t}_\infty, \tag{44}$$

for any $m \in \mathcal{H}_K$. Extending by continuity this relation to all $t \in \mathcal{I}$, and using Eq. (19), the dense-monitoring limit of this linear functional defines Loève's isometry

$$\langle X, m\rangle_K = \lim_{N\to\infty} \mathbf{X}^\dagger\mathbf{K}^{-1}\mathbf{m} = \psi_X^{-1}(m). \tag{45}$$

The spectral form of this *congruence* inner product is (Berlinet and Thomas-Agnan 2004)

$$\langle X, m\rangle_K = \sum_{j=1}^{\infty} \frac{\mu_j \xi_j}{\lambda_j}, \tag{46}$$

where $\{\lambda_j\}_{j=1}^\infty$ are the eigenvalues of $\mathcal{K}$, and $\{\mu_j\}_{j=1}^\infty$, $\{\xi_j\}_{j=1}^\infty$ are the coefficients of the Karhunen-Loève expansions

$$m(t) = \sum_{j=1}^{\infty} \mu_j \phi_j(t), \tag{47}$$

$$X(t) = \sum_{j=1}^{\infty} \xi_j \phi_j(t), \tag{48}$$

respectively.

## 6.2 Singular (near perfect) homoscedastic classification

When $m \notin \mathcal{H}_K$, the measures $\mathbb{P}_0$ and $\mathbb{P}_1$ are mutually singular (orthogonal). In this case, near perfect classification is obtained (Parzen 1961a; Kailath 1966, 1971; Berrendero et al. 2018b). The terms $\langle X, m\rangle_K$ and $\langle m, m\rangle_K$ diverge. These divergences, which are of the same type, dominate in Eq. (39). Therefore the term that depends on the class priors, which is non-singular for $0 < p < 1$, can be dropped out. With an appropriate interpretation of the limit, the decision rule is

$$\mathbb{I}\left[\lim_{\hat{m}_\mathcal{H}\to m}\left(\langle X, \hat{m}_\mathcal{H}\rangle_\mathcal{H} - \frac{1}{2}\langle \hat{m}_\mathcal{H}, \hat{m}_\mathcal{H}\rangle_K\right) > 0\right], \tag{49}$$

where $\hat{m}_\mathcal{H} \in \mathcal{H}_K$ is an approximation to the mean $m \notin \mathcal{H}_K$, whose squared norm is $\langle \hat{m}_\mathcal{H}, \hat{m}_\mathcal{H}\rangle_K$. The quantity $\langle X, \hat{m}_\mathcal{H}\rangle_K$ is defined through Loève's isometry.

The limit in Eq. (49) needs to be understood as follows: Since the elements of $\mathcal{H}_K$ are dense in $L^2(\mathcal{I})$ when $\mathcal{K}$ has no zero eigenvalues (Cucker and Zhou 2007), it is possible to build a sequence of approximating classification problems with $\hat{m}_\mathcal{H} \in \mathcal{H}_K$ that converges to $m \in L^2(\mathcal{I})$. The singular homoscedastic classification problem with $m \notin \mathcal{H}_K$ can be seen as the limit of a sequence of classification problems involving functions $\hat{m}_\mathcal{H} \in \mathcal{H}_K$ in the approximating sequence (Theorem 6 of Berrendero et al. (2018b)). An optimal classification rule for these related problems is given by Eq. (39). Since $\langle m, m\rangle_K$ diverges, the corresponding classification errors, which are given by limit of Eq. (40), tend to zero.

### *6.2.1 Brownian processes with different means*

This singular limit can be illustrated in the discrimination of trajectories sampled from one of two Brownian processes with the same variance but different means. Let us consider a homoscedastic classification problem in which the class 0 trajectories are realizations of a zero-mean Brownian process in $t \in [0, T]$ and the class 1 trajectories are sampled from a Brownian process with a piecewise linear mean

$$m(t) = \begin{cases} 0 & 0 \le t < t_1 \\ m_T \frac{t-t_1}{t_2-t_1} & t_1 \le t < t_2 \\ m_T & t_2 \le t < T \end{cases}, \tag{50}$$

with $m_T = m(T)$, a constant, and $0 < t_1 \le t_2 < T$.

The Brownian process kernel is

$$K_{BM}(s,t) = \sigma\,\min\{s,t\}, \quad \sigma > 0. \tag{51}$$

The RKHS associated with this kernel is $\mathcal{H}_{BM}$, the Sobolev space of absolutely continuous functions $f$ in $t \in [0,T]$, such that $f(0) = 0$, and whose derivatives are square integrable in that time interval (i.e. $f' \in L^2[0,T]$). The corresponding inner product between $f, g \in \mathcal{H}_{BM}$ is

$$\langle f, g\rangle_{BM} = \frac{1}{\sigma^2}\int_0^T f'(t)g'(t)dt. \tag{52}$$

The mean $m$ given by Eq. (50) is in $\mathcal{H}_{BM}$ provided that $t_1 < t_2$. In such case, its squared norm is

$$\langle m, m\rangle_{BM} = \frac{1}{\sigma^2}\int_0^T |m'(t)|^2\,dt = \frac{m_T^2}{\sigma^2}\frac{1}{t_2 - t_1}. \tag{53}$$

Since $\mathcal{H}_{BM}$ is an infinite-dimensional RKHS, the sample trajectories $X$ do not belong to that space with probability one (Lukić and Beder 2001). In the case of Brownian motion it is clear that the sample trajectories, which are continuous but non-differentiable, are not in $\mathcal{H}_{BM}$. In consequence, the term $\langle X, m\rangle_{BM}$, which appears in the

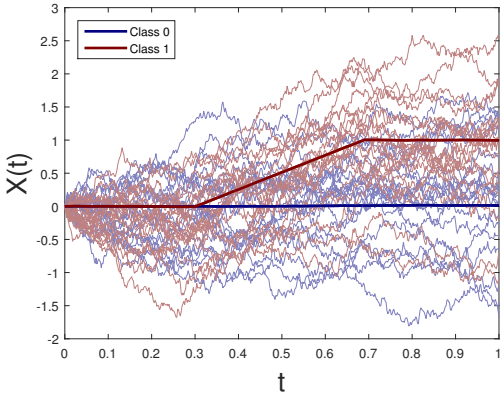


**Fig. 1** Homoscedastic classification: zero-mean Brownian motion vs. Brownian motion with a piecewise linear mean.

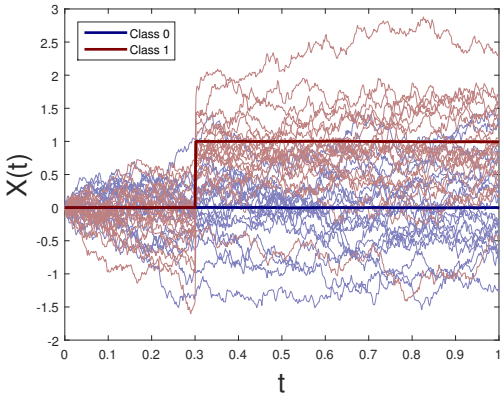


**Fig. 2** Homoscedastic near perfect classification: zero-mean Brownian motion vs. Brownian motion with a step function mean.

decision rule, cannot be understood as an inner product in that space. Nonetheless, it can be computed starting from Eq. (52) and integrating by parts

$$\langle X, m\rangle_{BM} = \int_0^T X'(t)m'(t)dt = -\int_0^T X(t)m''(t)dt = \frac{m_T}{\sigma^2}\frac{X(t_2)-X(t_1)}{t_2-t_1}. \quad (54)$$

To derive this expression we have employed the fact that the Brownian trajectories start at the origin ($X(0)=0$), and that for the piecewise linear mean given by Eq. (50), $m'(T)=0$. Using Eqs. (39), (53), and (54), the Bayes rule becomes

$$\mathbb{I}\left[\frac{m_T}{\sigma^2}\left((X(t_2)-X(t_1))-\frac{m_T}{2}\right) > (t_2-t_1)\log\frac{1-p}{p}\right].$$

A non-singular classification problem of this type is depicted in Fig. 1, where $T=1$, $m_T=1$, $t_1=0.3$, $t_2=0.7$, and $\sigma=1$.

In the limit $t_2 \to t_1^+$, the mean exhibits a finite discontinuity at $t_1$ and therefore, is not in $\mathcal{H}_{BM}$. In this case, the decision rule is

$$\mathbb{I}\left[\left(X(t_1^+)-X(t_1)\right) > \frac{m_T}{2}\right], \quad (55)$$

and near perfect classification (zero asymptotic error) is obtained. As is apparent from Fig. 2, this rule has an obvious interpretation: one needs to compare the values of trajectory immediately before and after the jump of $m(t)$ at $t=t_1$. Class 0 trajectories should be continuous. Class 1 trajectories should exhibit the same discontinuity as the mean. This rule guarantees perfect classification provided that the values of the trajectories can be monitored with arbitrarily high resolution in $t$.

## 7 Heteroscedastic classification problems

In contrast with the homoscedastic case, which has received wide attention in the literature (Parzen 1961a; Kailath 1966, 1971; Delaigle and Hall 2012; Berrendero et al. 2018b), most work on the heteroscedastic case is fairly recent and limited to specific examples (Delaigle and Hall 2012, 2013; Dai et al. 2017; Berrendero et al. 2018b). In fact, no general rule has been proposed in the literature for this setting, even in the non-singular case. The difficulty lies in the interpretation of the singular terms that appear in the optimal rule (Eq. (28)). As discussed in Section 5, when the divergences cancel out, we have a non-singular classification problem. By contrast, if the divergences do not cancel out, an optimal decision rule can be derived by balancing the singular terms. In this singular case near perfect classification is obtained. We shall now proceed to study these cases separately and in detail.

### 7.1 Non-singular heteroscedastic classification

A heteroscedastic classification problem is non-singular if the Gaussian process laws $\mathbb{P}_0$ and $\mathbb{P}_1$ are equivalent. For this to be the case, $m \in \mathcal{H}_0 \cap \mathcal{H}_1$, so that the terms that involve $m$ in Eq. (28) must be well-defined. Furthermore, the singularities of the quadratic terms in the optimal rule need to cancel out: On the one hand, the limit

$$\lim_{N\to\infty} \mathbf{X}^\dagger\left(\mathbf{K}_1^{-1}-\mathbf{K}_0^{-1}\right)\mathbf{X} \equiv \langle X,X\rangle_{K_1} - \langle X,X\rangle_{K_0} \quad (56)$$

should be finite when the set of monitoring points becomes dense in $\mathcal{I}$. On the other hand, the limit

$$\lim_{N\to\infty}\prod_{j=1}^{N}\frac{\lambda_{1j}}{\lambda_{0j}} \equiv \frac{|\mathcal{K}_1|}{|\mathcal{K}_0|}. \quad (57)$$

should exist and be different from zero.

If these conditions are obtained, the divergences in the individual terms of Eq. (28) cancel out and the resulting classification rule is well defined. Formally, it can be written as

$$\mathbb{I}\Bigg[ -\frac{1}{2}\log\frac{|\mathcal{K}_1|}{|\mathcal{K}_0|} \tag{58}$$
$$-\frac{1}{2}\left(\langle X-m, X-m\rangle_{K_1} - \langle X, X\rangle_{K_0}\right) > \log\frac{1-p}{p}\Bigg].$$

In the following subsection, this non-singular limit will be illustrated using the standard Brownian motion and the standard Brownian bridge processes in the interval $[0,T]$, which are known to be equivalent when $0 \leq T < 1$ (Varberg 1961; Shepp 1966). Therefore, for this range of values of $T$, the problem is heteroscedastic, but not singular. It becomes singular at $T = 1$.

*7.1.1 Standard Brownian vs. Brownian bridge processes*

The standard Brownian bridge in $[0,1]$ is a zero-mean Gaussian process whose kernel is

$$K_{BB}(s,t) = \min\{s,t\} - st, \quad s,t \in [0,1]. \tag{59}$$

The corresponding RKHS is

$$\mathcal{H}_{BB} = \left\{ f : f(t) = \int_0^t f'(s)ds; f(1) = 0; f' \in L^2[0,1] \right\}.$$

This process, if considered in the interval $[0,T]$, with $T < 1$, has the inner product

$$\langle f, g\rangle_{BB} = \int_0^T f'(t)g'(t)dt + \frac{f(T)g(T)}{1-T} \tag{60}$$

In this interval, the Brownian bridge is equivalent to the standard Brownian process, whose kernel is

$$K_{BM}(s,t) = \min\{s,t\}, \tag{61}$$

and whose associated RKHS is

$$\mathcal{H}_{BM} = \left\{ f : f(t) = \int_0^t f'(s)ds; f' \in L^2[0,1] \right\}. \tag{62}$$

When restricted to the interval $[0,T]$, its inner product is

$$\langle f, g\rangle_{BM} = \int_0^T f'(t)g'(t)dt. \tag{63}$$

Let $X$ be a trajectory in the $[0,T]$ interval that is either a sample from a Brownian motion process, with probability $1-p$ (class 0), or from a Brownian bridge process, with probability $p$ (class 1). Trajectories from either of these processes are continuous but not differentiable. Therefore, they do not belong to the corresponding Hilbert spaces. In consequence, the individual inner products $\langle X, X\rangle_{K_i}$, for $i \in \{0,1\}$ are singular. However, the difference

$$\langle X, X\rangle_{BB} - \langle X, X\rangle_{BM} = \frac{(X(T))^2}{1-T}, \tag{64}$$

is well defined for $0 \leq T < 1$ because the singular terms in Eqs. (60) and (63), which involve the derivatives $f'$ and $g'$, are identical, and therefore cancel out.

To derive the expression for the ratio of determinants in Eq. (58) we consider the discretely monitored process in $[\frac{1}{N}, T]$ at regularly spaced times $\{t_n = n\Delta T\}_{n=1}^{T}$, with $\Delta T = \frac{T}{N}$ for some integer $N$, which will eventually be made to approach $\infty$. The point $t_0 = 0$ is excluded because both processes take the same deterministic value (i.e., $X(t=0) = 0$).

The Gram (autocovariance) matrix of such a discretely monitored standard Brownian process is

$$(\mathbf{K}_{BM})_{mn} = \Delta T \min\{m,n\}, \quad m,n = 1,\ldots,N. \tag{65}$$

The determinant of this matrix is

$$|\mathbf{K}_{BM}| = (\Delta T)^N. \tag{66}$$

The corresponding Gram matrix for the discretely monitored standard Brownian bridge is

$$(\mathbf{K}_{BB})_{mn} = \frac{T}{N}\left(\min\{m,n\} - mn\frac{T}{N}\right), \tag{67}$$

for $m,n = 1,\ldots,N$. The determinant of this matrix is

$$|\mathbf{K}_{BB}| = (1-T)(\Delta T)^N. \tag{68}$$

Thus, the ratio of the determinants of the covariance matrices for the discretely monitored standard Brownian motion and the standard Brownian bridge is

$$\frac{|\mathbf{K}_1|}{|\mathbf{K}_0|} = 1 - T, \quad \forall N > 0. \tag{69}$$

Therefore,

$$\frac{|\mathcal{K}_{BB}|}{|\mathcal{K}_{BM}|} = \lim_{N\to\infty}\frac{|\mathbf{K}_{BB}|}{|\mathbf{K}_{BM}|} = 1 - T. \tag{70}$$

Using this result in Eq. (28), we get the optimal classification rule for this problem (see, e.g., Berrendero et al. (2018b))

$$\mathbb{I}\left[ -\frac{1}{2}\log(1-T) - \frac{1}{2}\frac{(X(T))^2}{1-T} > \log\frac{1-p}{p}\right]. \tag{71}$$

The error of this rule is

$$\begin{aligned} L^* =& (1-p)\left(1 - 2\Phi\left(\frac{-D}{\sqrt{T}}\right)\right) \\ &+ p\left(2\Phi\left(\frac{-D}{\sqrt{T(1-T)}}\right)\right), \end{aligned} \tag{72}$$

where $\Phi$ is the cumulative distribution function of a standard normal distribution, and

$$D = \sqrt{-2(1-T)\left(\log\left(\frac{1-p}{p}\right) + \frac{1}{2}\log(1-T)\right)}.$$

Note that in the limit $T \to 1^-$, the two terms on the left of the expression within the indicator function diverge, and dominate the classification rule. These divergences signal that near perfect classification classification is obtained in this limit. Dropping the term that involves the priors, which is not singular, the optimal rule in the limit $T \to 1^-$ becomes

$$\mathbb{I}\left[(X(T))^2 < (1-T)\log\frac{1}{1-T}\right]. \tag{73}$$

Since $(X(T))^2 \geq 0$ and the term on the right hand side approaches zero, the optimal rule for $T = 1$ is

$$\mathbb{I}\left[X(1) = 0\right]. \tag{74}$$

That is, one needs to inspect the value of $X(1)$. This quantity is 0 for Brownian bridge trajectories, and different from 0 with probability 1 for Brownian trajectories.

Finally, using result given by Eq. (36) of Section 5, the Radon-Nikodym derivative of the Brownian bridge measure with respect to the Brownian motion measure is

$$\frac{d\mathbb{P}_{BB}}{d\mathbb{P}_{BM}}(x) = \left(\frac{1}{1-T}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{(x(T))^2}{1-T}\right\}. \tag{75}$$

It is straightforward to also verify that the inner-products of the two processes are related by Eq. (37) with $\delta K(s,t) = -st$

$$\begin{aligned} &-\langle f, \langle g, \delta K\rangle_{BB}\rangle_{BM} \\ &\quad = \int_0^T f'(t)\,\frac{d}{dt}\left[\int_0^T g'(s)\,t\,ds + \frac{g(T)\,Tt}{1-T}\right]dt \\ &\quad = f(T)\left[g(T) + \frac{g(T)T}{1-T}\right] = \frac{f(T)g(T)}{1-T} \\ &\quad = \langle f, g\rangle_{BB} - \langle f, g\rangle_{BM}. \end{aligned}$$

## 7.2 Singular (near perfect) heteroscedastic classification

In the heteroscedastic setting a first type of singular classification problem arises when $m = m_1 - m_0 \notin \mathcal{H}_0 \cap \mathcal{H}_1$ (Delaigle and Hall 2012). In this case, the analysis made in Section 6.2 remains valid and near perfect classification is obtained. For this case, an optimal classification rule is (49) with $K = K_1$.

A second mechanism for near perfect classification is obtained if the singularities in the terms $\log\frac{|\mathcal{K}_1|}{|\mathcal{K}_0|}$ and $(\langle X, X\rangle_{K_0} - \langle X, X\rangle_{K_1})$, do not separately cancel out. In this case, the measures induced by $GP(0, K_0)$ and $GP(m, K_1)$ are mutually singular. The decision rule Eq. (28) is dominated by the divergent terms. One can therefore drop the terms that involve $m$ and the class priors, which are non-singular, and obtain the near perfect classification rule

$$\mathbb{I}\left[\frac{\langle X, X\rangle_{K_0} - \langle X, X\rangle_{K_1}}{\log|\mathcal{K}_1| - \log|\mathcal{K}_0|} > 1\right]. \tag{76}$$

In this rule, the ratio of divergent terms needs to be understood as

$$\mathbb{I}\left[\lim_{N\to\infty}\frac{\mathbf{X}^\dagger\left(\mathbf{K}_0^{-1} - \mathbf{K}_1^{-1}\right)\mathbf{X}}{\log|\mathbf{K}_1| - \log|\mathbf{K}_0|} > 1\right], \tag{77}$$

in the limit of dense monitoring.

In what follows, the validity of Eq. (77) is illustrated in the classification of two Brownian processes with equal mean and different variances, which are known to be mutually singular.

### *7.2.1 Classification of Brownian processes with different variances*

Consider the heteroscedastic functional classification problem

$$X(t) = \begin{cases} Z_0(t) & \text{w. p.} \quad 1-p \\ Z_1(t) & \text{w. p.} \quad p \end{cases}, \tag{78}$$

for $t \in [0, T]$, where $Z_0(t)$ and $Z_1(t)$ are zero-mean Brownian processes of variances $\sigma_0^2$ and $\sigma_1^2$, respectively. Since all trajectories start at the same level $X(0) = 0$, they need to be monitored only at times $\{t_n = n\Delta T\}_{n=1}^N$ with $\Delta T = T/N$. The autocovariance matrices of the discretely monitored processes are

$$(\mathbf{K}_i)_{mn} = \sigma_i^2 \Delta T \min\{m, n\}, \quad m, n = 1, \ldots, N, \tag{79}$$

for $i = 0, 1$. The determinant of this matrix is

$$|\mathbf{K}_i| = \left(\sigma_i^2 \Delta T\right)^N. \tag{80}$$

The corresponding inverses are symmetric tridiagonal matrices

$$\mathbf{K}_i^{-1} = \frac{1}{\sigma_i^2 \Delta T}\begin{pmatrix} 2 & -1 & 0 \ldots & 0 & 0 \\ -1 & 2 & -1 \ldots & 0 & 0 \\ 0 & -1 & 2 \ldots & 0 & 0 \\ & & \cdots & & \\ 0 & 0 & 0 \ldots & 2 & -1 \\ 0 & 0 & 0 \ldots & -1 & 1 \end{pmatrix}. \tag{81}$$

The term in the denominator of Eq. (77) is

$$\log|\mathbf{K}_1| - \log|\mathbf{K}_0| = \sum_{j=1}^N \log\frac{\sigma_1^2}{\sigma_0^2} = \frac{T}{\Delta T}\log\frac{\sigma_1^2}{\sigma_0^2}, \tag{82}$$

where we have used that $N = \frac{T}{\Delta T}$. Similarly,

$$\mathbf{X}^\dagger \mathbf{K}_i^{-1} \mathbf{X} = \frac{T}{\Delta T} \frac{\sigma_X^2(N)}{\sigma_i^2}, \tag{83}$$

where

$$\sigma_X^2(N) = \frac{1}{N} \sum_{n=1}^{N} \frac{\left(X(t_n) - X(t_{n-1})\right)^2}{\Delta T}. \tag{84}$$

For this problem, if the non-singular terms are dropped, Eq. (27) becomes

$$\mathbb{I}\left[\sigma_X^2(N) > \theta\right], \tag{85}$$

with $\theta = \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)^{-1} \log \frac{\sigma_1^2}{\sigma_0^2}$. In the limit $N \to \infty$ (therefore, $\Delta T \to 0^+$), the optimal classification rule is

$$\mathbb{I}\left[\sigma_X^2 > \theta\right], \tag{86}$$

where $\sigma_X^2 = \lim_{N\to\infty} \sigma_X^2(N)$. This near perfect classification rule can be written also in terms of Kullback-Leibler divergences between normal distributions with the same mean and different variances

$$\mathbb{I}\left[KL\left(N(0,\sigma_X), N(0,\sigma_1)\right) < KL\left(N(0,\sigma_X), N(0,\sigma_0)\right)\right]. \tag{87}$$

As in the homoscedastic case, this rule guarantees perfect classification only if the values of the trajectories can be measured with arbitrarily high resolution in time.

We will now analyze the convergence of the singular decision rule for the discretely monitored Brownian processes (Eq. (85)) to its asymptotic limit (Eq. (86)). Without loss of generality, we will assume $\sigma_1^2 > \sigma_0^2$. The accuracy of the prediction rule given by Eq. (85) as a function of $N$, the number of monitoring points, is

$$Acc(N) = pP\left(\sigma_X^2(N) > \theta \mid Y = 1\right) + (1-p)\, P\left(\sigma_X^2(N) < \theta \mid Y = 0\right). \tag{88}$$

Since $\left(X(t_n) - X(t_{n-1})\right) \sim N\left(0, \sigma_X\sqrt{\Delta T}\right)$, then $N\frac{\sigma_X^2}{\sigma_i^2}$ follows a chi-square distribution with $N$ degrees of freedom. In consequence,

$$Acc(N) = p\left[1 - \mathrm{cdf}_{\chi^2}\left(N\frac{\theta}{\sigma_1^2}, N\right)\right] + (1-p)\,\mathrm{cdf}_{\chi^2}\left(N\frac{\theta}{\sigma_0^2}, N\right), \tag{89}$$

where $\mathrm{cdf}_{\chi^2}(z,\nu)$ is the cumulative distribution function of a $\chi^2$ distribution with $\nu$ degrees of freedom evaluated at $z$. In the asymptotic limit of a densely monitored process $\lim_{N\to\infty} Acc(N) = 1$, which means that near perfect classification is obtained.

To illustrate this convergence, we have performed a set of experiments with simulated trajectories of zero-mean Brownian processes of variances $\sigma_0^2 = 1$ and $\sigma_1^2 \in$

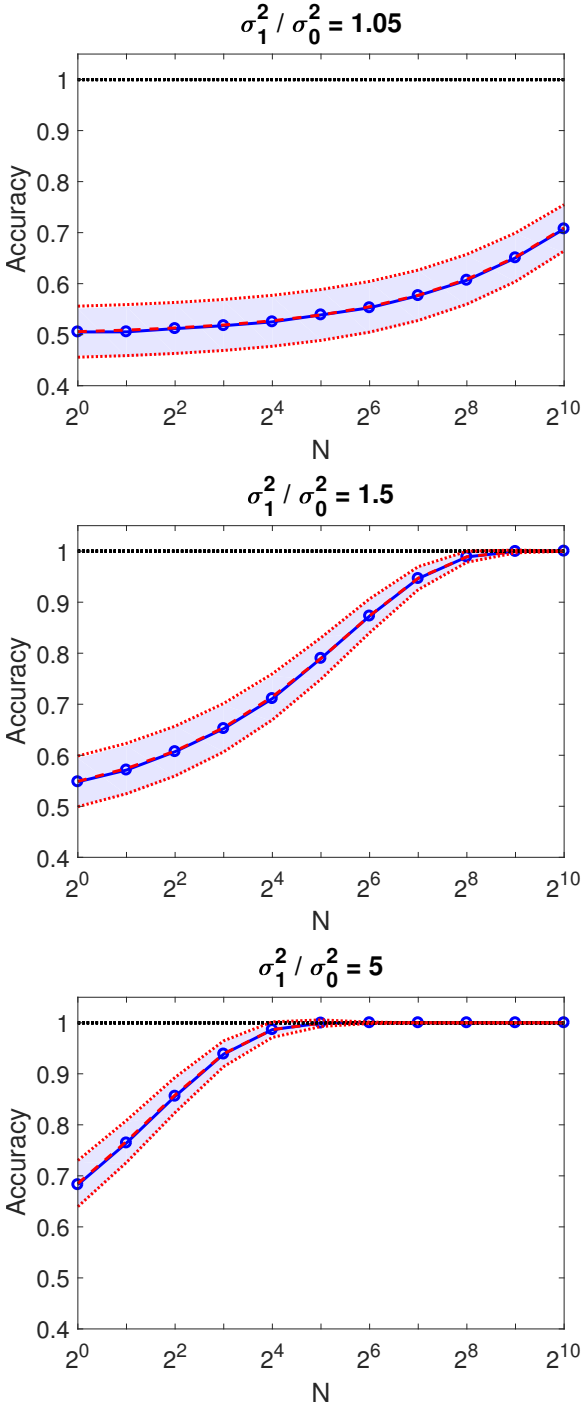


**Fig. 3** Classification accuracy for the discrimination of Brownian processes with equal means and different variances (Eq. 78). The solid lines trace the dependence of the accuracy, averaged over 1000 replications of the problem, as a function of $N$, the number of monitoring intervals. The shaded band corresponds to one standard deviation above and below this average. The dashed lines correspond to the theoretical expected accuracy. Finally, the red dotted lines correspond to the expected accuracy plus/minus one standard deviation.

$\{1.05, 1.5, 5\}$ in the time interval $[0, 1]$, starting from 0 at $t = 0$. Each experiment consists in generating $M = 50$ trajectories from each of these classes. The trajectories are sampled at $N+1$ regularly spaced times, including the origin, with $N = 2^b$, $0 \leq b \leq 10$. Then, the decision rule given by Eq. (85) is used to classify the $2M$ trajectories generated. The whole process was repeated 1000 times so that, for each $N$, the expected accuracy and its standard deviation of the accuracy on the sample of $2M = 100$ trajectories can be computed. In a sample of $2M$ trajectories the number of correctly classified cases follows a binomial distribution with a success probability equal to $Acc(N)$. Since the variance of the number of successes in the binomial distribution is given by $2MAcc(N)\,(1 - Acc(N))$, the standard deviation of the observed accuracy of the decision rule given by Eq. (85) in a sample of size $2M$ is $\sqrt{\frac{Acc(N)\cdot(1-Acc(N))}{2M}}$ with $Acc(N)$ given by Eq. (89).

Fig. 3 displays the dependence of the accuracy of the optimal decision rule as a function of $N$, for the different values of the ratio $\sigma_1^2/\sigma_0^2$ considered. From the analysis of these plots one concludes that sample estimates are in good agreement with the theoretical values of the expected accuracy and their standard deviations. Furthermore, it is apparent that the larger the differences between $\sigma_1^2$ and $\sigma_0^2$ are, the faster the approach to the asymptotic regime in which near perfect classification is obtained.

## 8 Empirical evaluation

In this section we compare the performance of the limit rules derived in this work with functional classifiers proposed in the literature. As test-bed for comparison we use simulated data and a real-world problem from quantitative finance. To make it possible to reproduce the results, the code for the experiments is available at `https://github.com/GAA-UAM/GP-Bayes-Rules-Experiments`. The simulated data correspond to the classification problems considered in Subsections 6.2.1 (homoscedastic and singular), 7.1.1 (heteroscedastic and equivalent), and 7.2.1 (heteroscedastic and singular). Assuming that the classes are balanced ($p = 1 - p = 1/2$), we generate $N_{train}$ trajectories in the interval $[0, T]$ and their corresponding class labels, $\mathcal{D}_{train} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_{train}}$, according to Eq. (3). To investigate how the accuracies of the different methods depend on the amount of data available for induction, experiments with different training set sizes ($N_{train} \in \{50, 200, 1000\}$) have been carried out. The trajectories $\{x(t),\, t \in [0, T]\}$ are monitored discretely on a regular grid

$$x(t_n), \quad t_n = \Delta T_b, \quad n = 0, 1, \ldots, N_b, \tag{90}$$

where $\Delta T_b = \frac{T}{N_b}$. The dependence on the size of the monitoring grid is analyzed by considering different numbers of discretization intervals; namely $N_b = 2^b$, $b = 1, \ldots, 10$. Unbiased estimates of the generalization accuracy are made in test sets of size $N_{test} = 1000$, which are generated independently of the training data. The values reported are averages over 100 independent replications, with the corresponding standard deviations.

The financial classification problem consists in the discrimination between the stocks of different car manufacturing companies (*BMW*, *GM*, and *Tesla*) on the basis of the time series of their market prices. According to expert knowledge, in this real-world example, the log-differences of the asset prices are expected to approximately follow a Brownian process (Osborne 1959; Fama 1965). The standard deviations of these processes, or, in financial terminology, their volatilities, should be different. Therefore, an appropriate model for their classification is given by (78); i.e., two Brownian processes with different variances. In this case, besides the comparison of methods, the experiments serve also as an empirical validation of this Brownian hypothesis.

The classifiers that are compared in this section are the following:

- LDA: The standard multivariate linear discriminant applied to the discretely monitored trajectories.
- QDA: The standard multivariate quadratic discriminant applied to the discretely monitored trajectories.
- PLS+Centroid: This classifier consists in applying a centroid rule to the output of a partial least squares regression model. It is one of the most accurate methods among those considered in the seminal paper by Delaigle and Hall (2012).
- PCA+QDA: This classifier is based on a proposal by Galeano et al. (2015) to compute the functional analogue of the Mahalanobis distance. In that paper, the authors argue that this method is equivalent to applying a quadratic discriminant to the first few principal components of the trajectories to be classified.
- RKC: The Reproducing Kernel classification rule is based on first performing variable selection according to a criterion that involves the Mahalanobis distance and then applying a linear discriminant analysis (Berrendero et al. 2018b). The name reflects the fact that it has a natural interpretation in the corresponding RKHS. This rule has been proven to be optimal if the functional classification problem is homoscedastic and the probability measures are equivalent.
- Limit-Rule: Classification rule derived from the analysis of the quadratic discriminant for the discretized process (Eq. (27)) in the limit of dense monitoring.

The different methods have been implemented in Python. LDA, QDA, PCA and PLS regression make extensive use of objects and functions in the *scikit-learn* package (Pedregosa et al. 2011). The RKC method has been freshly implemented following Berrendero et al. (2018b). Functional data objects have been manipulated with the tools provided by the *scikit-fda* package (Ramos-Carreño et al. 2019). The number of components of the dimensionality reduction methods (PCA, PLS and RKC) is determined by 10-fold cross-validation in the range 1 to 20.

We now proceed to present a summary of the results of this empirical evaluation for the different cases analyzed in this work.

### 8.1 Brownian processes with different means

We first study a homoscedastic problem of the form given by Eq. (3), in which $Z_0(t) = Z_1(t) = Z(t)$ is a standard Brownian

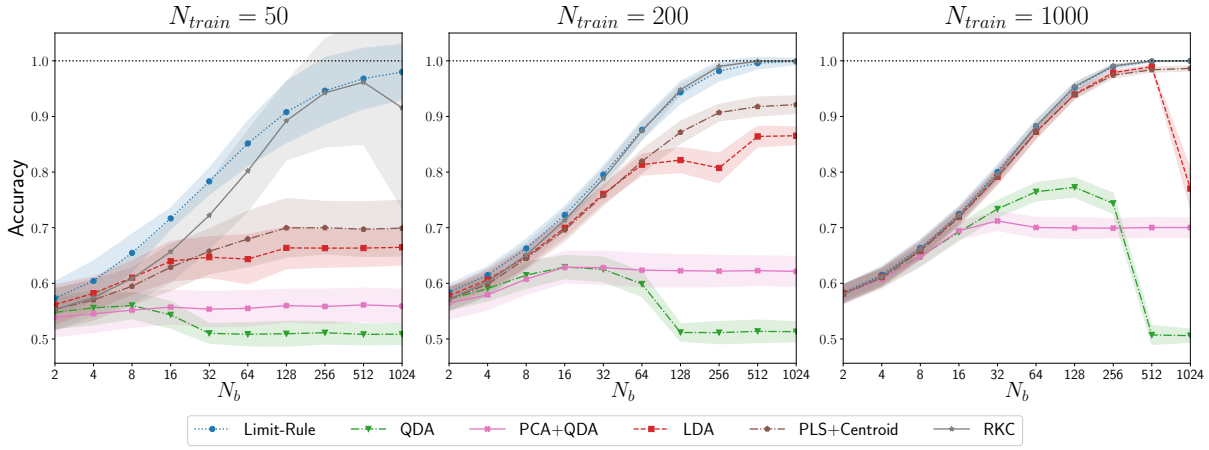


**Fig. 4** Classification accuracy for the discrimination of Brownian processes with equal variance and different means (zero mean, and step function mean). The solid lines trace the dependence of the accuracy, averaged over 100 replications of the problem, as a function of $N_b$, the number of monitoring intervals, for different values of $N_{train}$, the size of the training data. The shaded bands correspond to one standard deviation above and below the corresponding averages.

Motion process in $[0, 1]$ and $m(t)$ a step function

$$m(t) = \begin{cases} 0, & 0 \leq t \leq t_* \\ m_T, & t_* < t \leq 1 \end{cases}, \tag{91}$$

with $m_T$ a constant level. This corresponds to the problem analyzed in Subsection 6.2.1, with $t_2 \to t_1^+$, $t_1 = t_*$ in Eq. (50), and $T = 1$. In such limit, the probability measures of the two Gaussian processes are mutually singular and near perfect classification is obtained. In the experiments carried out, the step is located at $t_* = 0.5$ and has a height of $m_T = 0.3$.

The limit-rule for this problem is given by Eq. (55). It depends only on the location and the size of the discontinuity. For this rule, $t_*$, the time instant at which the discontinuity occurs, is estimated from the training data. Specifically, a two sided t-test is used to determine whether the difference of the sample means in class 1 trajectories between consecutive monitoring points are significantly different from zero. The discontinous jump is assumed to be within the interval $\left[\hat{t}_*, \hat{t}_* + \Delta T_b\right]$, where $\left\{\hat{t}_*, \hat{t}_* + \Delta T_b\right\}$ is the pair of consecutive points for which this test yields the lowest p-value. The height of the step $m_T$ is estimated as the empirical mean of the values of the class 1 trajectories right after the step

$$\hat{m}_T = \frac{1}{N_{train}^{[1]}} \sum_{i=1}^{N_{train}} x_i(\hat{t}_* + \Delta T_b)\mathbb{I}[y_i = 1], \tag{92}$$

where $N_{train}^{[1]} = \sum_{i=1}^{N_{train}} \mathbb{I}[y_i = 1]$ is the number of class 1 trajectories in the training set.

The curves plotted in Fig. 4 display the dependence of the average accuracies of the different classifiers as a function of the number of discretization intervals for different training set sizes. The shaded bands correspond to deviations of one standard deviation above and below the average accuracies. The black horizontal dotted line marks the optimal accuracy, which in this case is 1.0. From the results obtained we observe that the limit rule approaches this value asymptotically, for sufficiently dense monitoring. The RKC method performs remarkably well in this problem and also approaches perfect accuracy for large training samples and dense monitoring. For Brownian processes, the RKC method can be shown to be optimal for a difference of means that is a continuous piecewise linear function starting at 0 (Berrendero et al. 2018b). The problem considered in our simulation is not of this form, because the step function exhibits a discontinuity, albeit finite. Nevertheless, as discussed in Subsection 6.2.1, the discontinuous jump can be obtained as the limit of a sequence of such continuous functions. Therefore, it is reasonable that RKC performs as well as the limit rule, which is optimal.

The differences between the accuracies of these two methods and the remaining ones, which are small for coarse monitoring, become larger as $N_b$ increases. The superior performance of RKC and the limit rule also at small training sizes resides in the fact that they require estimating fewer parameters. In this problem, all the information needed for discrimination is in the difference of means between the two Brownian processes. For this reason, the classifiers that require the estimation of the covariance matrix, especially QDA, and PCA+QDA, which furthermore do not assume homoscedasticity, obtain very poor results. This is consistent with previous observations in the literature on the limited accuracy of quadratic discriminant functions when the dimensions are large and the sample sizes for the estimation of the covariances are small (Marks and Dunn 1974; Wahl and Kronmal 1977; Berrendero and Cárcamo 2019). For larger values of $N_b$, PCA+QDA, which involves a dimensionality reduction step before the quadratic determinant function is computed, becomes more accurate than standard QDA. This behavior is consistently observed for all the classification problems analyzed.

The accuracies of PLS+Centroid and LDA are also low when the training sets are small. Both are global methods, which are not well adapted to problems in which the discriminant information is concentrated at a single point. Nonetheless, their accuracy markedly improves (especially that of PLS+Centroid) as the size of the training data becomes larger.

Finally, note that even though QDA and LDA are optimal for the multivariate version of this problem, the collinearity inherent to functional data has a marked negative impact in their predictive performance for finite training sample sizes. The reason is that these methods require the inversion of the empirical covariance matrix. This inversion is numerically unstable when the number of variables (monitoring times) increases for a fixed size of the training sample. By contrast, the accuracies of PLS+Centroid, PCA+QDA, and RKC (which makes use of LDA) do not deteriorate with increasing $N_b$,

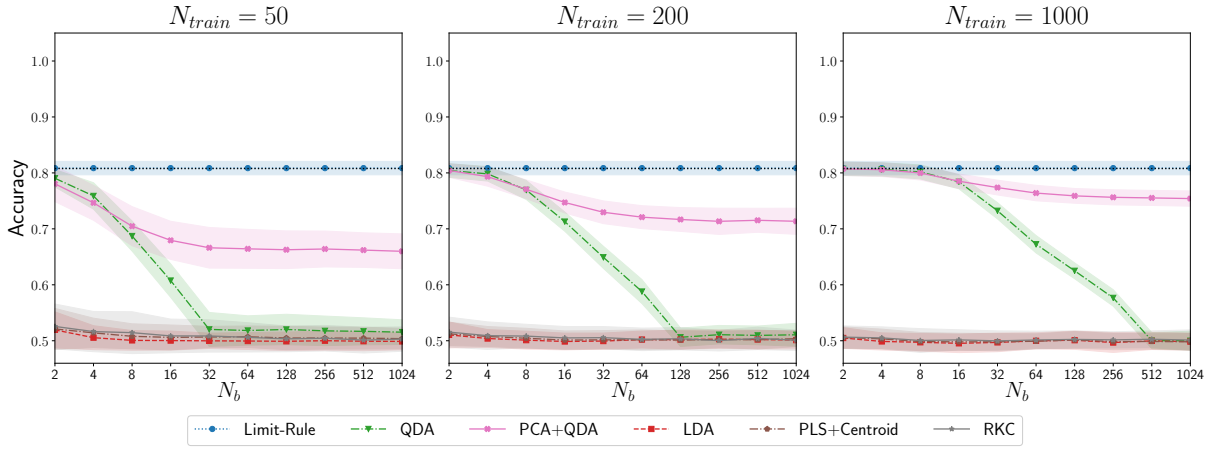


**Fig. 5** Classification accuracy for the discrimination of standard Brownian and Brownian bridge processes in $[0, T]$ with $T = 0.95$. The solid lines trace the dependence of the accuracy, averaged over 100 replications of the problem, as a function of $N_b$, the number of monitoring intervals, for different values of $N_{train}$, the size of the training data. The shaded bands correspond to one standard deviation above and below the corresponding averages.

because they involve a dimensionality reduction in a previous step.

## 8.2 Brownian motion vs. Brownian bridge

The problem addressed in this second batch of experiments is the discrimination of trajectories sampled from a standard Brownian process and from a Brownian bridge process in the interval $[0, T]$. As described in Section 7.1.1, these processes are equivalent for $T < 1$. In the experiments carried out the value selected is $T = 0.95$. Therefore, this is a standard classification problem with a non-zero Bayes error. Specifically, for the current simulation Eq. (72) yields $L^* = 0.193$. The limit rule is Eq. (71). It depends only on the class priors and on the value of the trajectory to be classified at $T$.

In Fig. 5 we present the comparison among the classifiers described in the introduction of the current section. As expected, the average accuracy obtained with the limit rule classifier is close to the optimal value of $1 - L^* = 0.807$ for all values of $N_{train}$ and $N_b$. In this case, since both processes have the same mean, the information that is useful for discrimination is in the covariance structure. Therefore linear classifiers, such as LDA, PLS+Centroid, and RKC are unable to predict better than random guessing. Both QDA and PCA+QDA obtain good results when the number of monitoring intervals is small and the size of the training data is large. Their predictive performance deteriorates as $N_b$ becomes larger. As in the previous case, the reason can be traced to the estimation of the covariance matrix from the sample, which becomes unreliable at higher dimensions. PCA+QDA is more robust than QDA because of the dimensionality reduction step.

## 8.3 Brownian processes with different variances

We now address the classification of trajectories sampled from two zero-mean Brownian processes of different variances. The problem, which has been analyzed in detail in Section 7.2.1, is singular and exhibits near perfect classification. The limit rule is given by Eq. (86). It requires the estimation of $\sigma_0^2$, $\sigma_1^2$ and $\sigma_X^2$ from the training data. The variance of each trajectory is estimated using Eq. (84). The variances $\sigma_0^2$ and $\sigma_1^2$ are estimated as the averages of the variances in the class 0 and class 1 trajectories, respectively. In the experiments performed the class 0 trajectories are realizations of a standard Brownian process with $\sigma_0^2 = 1$. The class 1 trajectories are sampled from a Brownian process with $\sigma_1^2 = 1.5$.

The overall comparison of the different classifiers considered in this study for this problem is presented in Fig. 6. As in the previous set of experiments, since the two processes have the same mean, this is a purely heteroscedastic classification problem. In consequence, the linear methods, such as LDA, PLS+Centroid, and RKC, which are based solely on the differences between means, are equivalent to random guessing.

The predictions of QDA and PCA+QDA are better than random and improve with the size of the training data. Nonetheless, both methods are suboptimal. In particular, the accuracy of QDA severely deteriorates with the number of monitoring points, because of the increased dimension of the problem and the high collinearity of the functional data.

In this singular case, the limit rule approaches perfect accuracy when the number of monitoring intervals is sufficiently large even for small training samples. The reason is that the estimation given by Eq. (84) approaches the exact value of the variance in the limit $N \to \infty$ for a single trajectory. Therefore, the classification rule achieves perfect accuracy asymptotically, in the limit of dense monitoring, independently of the size of the training data.

## 8.4 Near perfect classification of financial time series

We now provide an illustration of near perfect classification with real-world data. The goal is to discriminate between time series of market prices of financial assets. In our experiments, the daily closing prices of General Motors (GM) from the New York Stock Exchange (NYSE), Tesla from NASDAQ, and BMW from Xetra, between January 1, 2014 and January 31, 2018, are used. The data have been retrieved via the *Google Finance* API (https://finance.google.com). Days in which not all three asset price quotations were available have been discarded. A more sophisticated treatment of missing

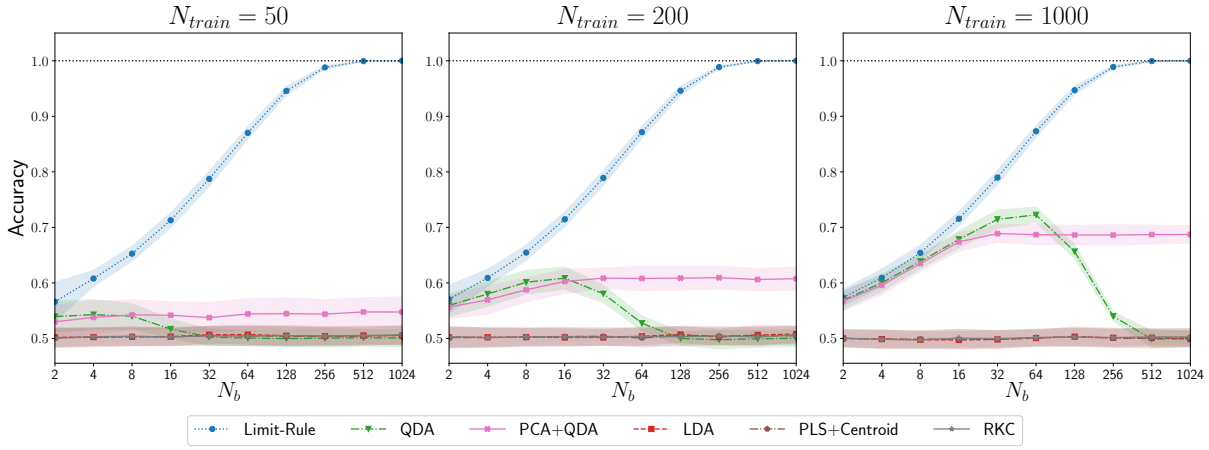


**Fig. 6** Classification accuracy for the discrimination of Brownian processes of different variances. The solid lines trace the dependence of the accuracy, averaged over 100 replications of the problem, as a function of $N_b$, the number of monitoring intervals, for different values of $N_{train}$, the size of the training data. The shaded bands correspond to one standard deviation above and below the corresponding averages.

values (e.g., linear or Brownian bridge interpolation) does not lead to significant changes of the results.

For each asset, the sample consists of $M = 30$ time series of market prices during non-overlapping periods of $N_B = 2^5 = 32$ days. The setup of the experiment is described in detail in Appendix B. These series are displayed in the top plots of Fig. 7.

There is ample empirical evidence that the time series of stock prices approximately follow a geometric Brownian process (Osborne 1959; Fama 1965). Consequently, their log-differences (i.e. the log-returns) follow an arithmetic Brownian process. According to standard financial wisdom, financial assets are characterized mainly by their *volatility*, which is the financial term used for the standard deviation of these log-returns. By contrast, the expected returns (i.e., the drift of the Brownian process) are much less reliable for discrimination. Therefore, Eq. (78), which corresponds to Brownian processes with different standard deviations (volatilities), should provide a suitable model for the classification problem. We can test the validity of these observations by comparing the accuracies of the classification methods described in the introduction to the current section and the limit rule given by Eq. (85). In this limit rule, the information on the means (expected returns) is discarded. Classification is made solely in terms of the sample estimates of the asset volatilities.

The results of the empirical evaluation are summarized in the bottom plots of Fig. 7. In each of the columns in this figure a different binary classification problem is considered. From left to right: BMW vs. GM, BMW vs. Tesla, and GM vs. Tesla. The inputs for classification are the discretely monitored trajectories of asset log-returns, which are computed as described in Appendix B. The accuracy of different classifiers is estimated using 10-fold stratified cross validation. The plots display the curves that trace the dependence of the accuracy of the different classifiers as a function of $N_b \in \{1, 2, 4, 8, 16, 32\}$, the number of monitoring intervals. The value $N_b = 32$ corresponds to daily intervals, which is the highest monitoring resolution that can be employed with the available data. For reference, we provide the theoretical accuracy curves for the corresponding Brownian processes with the same mean and volatility as each of the financial asset returns. Uncertainty intervals of one and two standard deviations above and below the theoretical accuracy curves are given as shaded bands.

In the first classification problem considered, BMW vs. GM, all classifiers perform poorly, close to random guessing. The reason is that these two assets have similar volatilities ($\sigma^2_{BMW} = 2.545 \cdot 10^{-4}$ and $\sigma^2_{GM} = 2.183 \cdot 10^{-4}$, respectively) and, in consequence, are difficult to distinguish. This should be expected because both companies are car manufactures that have comparable profiles and are exposed to the same risk factors. Therefore, the prices of their stock should exhibit similar characteristics. By contrast, Tesla is a highly specialized manufacturer of electric cars, whose main asset is technological innovation. Correspondingly, it exhibits higher volatility than the other two ($\sigma^2_{Tesla} = 6.147 \cdot 10^{-4}$). The characteristics of the *BMW vs. Tesla* and the *GM vs. Tesla* classification tasks are similar. In these two problems, the limit rule given by Eq. (85) has the best overall results. By contrast, the methods that rely on the difference of means (LDA, PLS+Centroid, and RKC) for discrimination have poor accuracies, at the level of random guessing. This means that the sample means (expected log-returns) are not useful to discriminate between these assets. The quadratic discriminant with a covariance matrix estimated from the sample (QDA) has slightly better accuracy than random guessing for intermediate values of $N_b$. However, for larger values of $N_b$ the results deteriorate. As in the synthetic data examples, this is a consequence of the poor quality of the sample estimates of the covariance matrices in higher dimensions and the high collinearity of functional data (Marks and Dunn 1974; Wahl and Kronmal 1977; Berrendero and Cárcamo 2019). The PCA+QDA method does not exhibit this degradation thanks to the fact that the dimension of the problem is reduced by selecting a few principal components before the quadratic discriminant rule is applied.

One way to avoid this limitation of the quadratic discriminant rule is to use expert knowledge and assume that the covariance matrix has a Brownian structure. Taking advantage of this structure, the elements of the covariance matrix need not be estimated separately. They can be computed in terms of the volatilities of the assets, which are the only parameters that are actually estimated from the training sample. To illustrate this point, the accuracy of this method (Brownian-QDA) is compared with the standard QDA, in which the

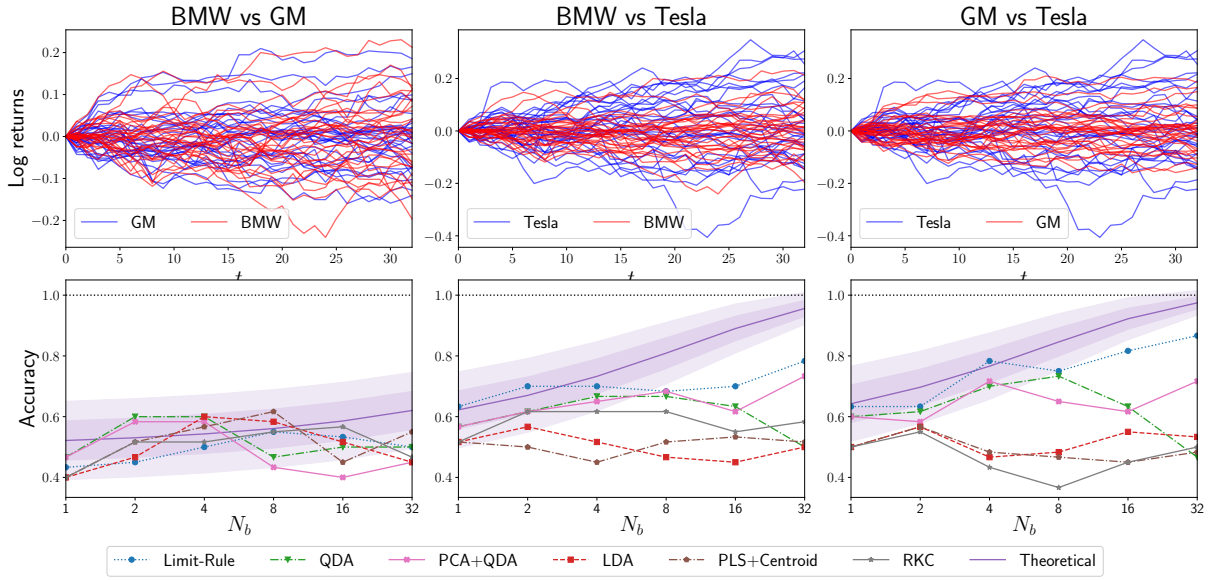


**Fig. 7** Discrimination of financial assets on the basis of the time series of their market prices (top plots). The curves that trace the dependence of the accuracies of different classifiers as a function of $N_b$, the number of monitoring intervals, are displayed in the bottom plots.

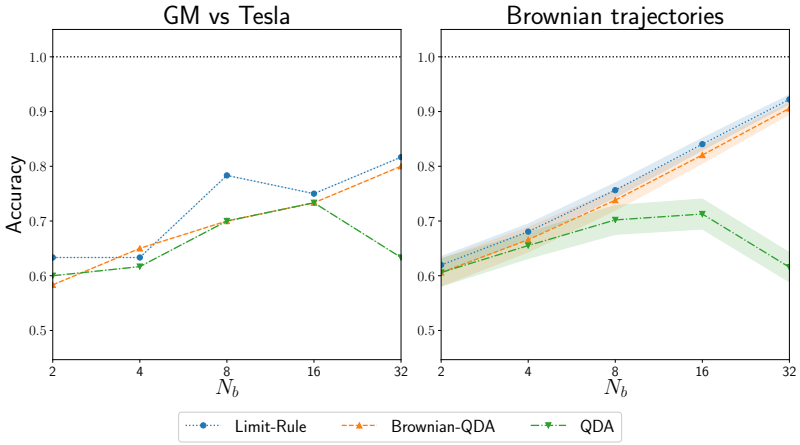


**Fig. 8** Comparison of QDA, Brownian QDA and the limit rule in the *GM vs. Tesla* classification problem (left plot). The right plot corresponds to the same comparison for simulated data of the same characteristics as the real-world problem.

individual elements of the covariance matrix are estimated separately, and the limit rule given by Eq. (85) in the GM vs. Tesla classification problem. The results of this comparison are displayed in the plot on the left-hand side of Fig. 8. By contrast with the behavior of the standard QDA, the accuracy of Brownian-QDA improves with the monitoring frequency. Nevertheless, comparable or better accuracies are achieved if we use the limit rule, in which only the singular terms in the quadratic discriminant are retained.

The time series of log-returns analyzed do not necessarily follow a Brownian process. Therefore, one may wonder whether the conclusions obtained with the real-world data are reliable. To clarify this point, we carried out simulations of the classification problem using trajectories from two Brownian processes with the same volatilities as the GM and Tesla assets. The results of these experiments are presented in the plot on the right-hand side of Fig. 8. To obtain these results the different classifiers (QDA, Brownian QDA, and the limit rule) are trained under the same conditions as in the experiments with the real-world data. Their accuracies are then computed on a test set of size 1000. The values reported are averages over 100 replications of the classification problem. Uncertainty intervals of one standard deviation above and below the averages are plotted as shaded bands. From these results one concludes that the behavior observed in the experiments with real-world data is not spurious: The predictive accuracy of the standard quadratic discriminant rule eventually deteriorates as $N_b$ increases. By contrast, the accuracies of Brownian-QDA and the limit rule given by Eq. (85) improve with denser monitoring. Note, however, that even for the largest values of $N_b$ considered, these classifiers do not achieve perfect classification. There are several reasons for this shortfall: First, the number of trajectories available for induction is very small (30 instances per class). Unfortunately, it is not possible to use much longer periods for which the hypothesis of constant volatility holds even in an approximate manner. Second, the daily monitoring

is insufficiently dense. However, higher frequency data cannot be used because the intra-day series of prices exhibits discontinuities and large deviations from the log-normal model. Finally, systematic deviations from the Brownian model are observed in the data. In the period considered, the Brownian assumption holds only in an approximate manner. Empirically, one observes that the time series exhibit heteroscedasticity in time and the log-returns are leptokurtic. Therefore, a more accurate model should account for the stochastic dynamics of the volatility (Bollerslev et al. 1992) and the heavy-tailedness of the log-returns (Cont 2001). In spite of these limitations, when the volatilities are sufficiently different, the limit rule given by Eq. (85), performs quite well in practice.

## 9 Conclusions

In this work we have addressed the problem of learning by induction from data that are characterized by functions of a continuous parameter. In particular, we have derived optimal classification rules for binary classification problems in which the instances are trajectories sampled from different Gaussian Processes, depending on the class label. The problem has been addressed earlier in the literature in both the homo- and heteroscedastic settings (see, e.g., Delaigle and Hall (2012, 2013); Dai et al. (2017); Berrendero et al. (2018b)). However, the procedure proposed in this work, which is based on the asymptotic analysis of the optimal rules for the discretely monitored trajectories in the limit of dense monitoring, is new. Furthermore, this procedure has been used to gain insight into the emergence of near perfect classification, which was first analyzed in Delaigle and Hall (2012) for differences in means. The current research expands on that work by analyzing cases in which near perfect classification arises from the covariance (quadratic) components as well. Specifically, a detailed analysis of the dense monitoring limit reveals that some of the terms that appear in such rules diverge. If the Gaussian processes are equivalent these divergences cancel out and non-singular optimal classification rules are obtained. By contrast, if the Gaussian processes are orthogonal the divergences do not cancel out. As a matter of fact, the singular terms dominate and near perfect classification is obtained. In this latter context, optimal rules that achieve zero prediction error asymptotically (i.e. for sufficiently large sample sizes) have been derived by considering only the terms that diverge in the limit of dense monitoring.

To illustrate the validity of the analysis, explicit rules are given for some classification problems involving Brownian and Brownian bridge processes. In the cases that such optimal rules were known, the limit of dense monitoring provides a novel procedure for their derivation. We also provide explicit rules for cases in which near perfect classification is obtained. The accuracy of such limit rules has been evaluated in extensive simulations and in the classification of time series of financial asset prices, which are modeled as geometric Brownian motion.

Even though the asymptotic analysis of the classification rules for the discretely monitored trajectories in the limit of dense monitoring has been introduced in the context of Gaussian process, the procedure may be applicable to more general stochastic processes, which are not necessarily Gaussian. This is a promising line of research that will be addressed in future work.

**Acknowledgements** The research has been supported by the Spanish *Ministry of Economy, Industry, and Competitiveness - State Research Agency*, projects MTM2016-78751-P and TIN2016-76406-P (AEI/FEDER, UE), and *Comunidad Autónoma de Madrid*, project S2017/BMD-3688. The authors gratefully acknowledge the use of the computational facilities at the *Centro de Computación Científica* (CCC) at the *Universidad Autónoma de Madrid* (UAM).

## A Discrete monitoring

In the derivations carried out the processes $X$ are monitored at a set of appropriately chosen discrete times $\{t_i\}_{i=i}^{N} \in \mathcal{I}^N$. The integrals that appear (e.g., in the definitions of the inner products) are then approximated by Riemann sums

$$\int_{t\in\mathcal{I}} h(t)dt \approx \frac{1}{N}\sum_{n=1}^{N} h(t_n). \tag{93}$$

For functions that are continuous in $\mathcal{I}$, these Riemman sums converge to the corresponding definite integrals in the limit of dense monitoring

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} h(t_n) = \int_{t\in\mathcal{I}} h(t)dt \quad \forall h \in \mathcal{C}\left[I\right]. \tag{94}$$

Let $K_0$ and $K_1$ be symmetric, strictly positive kernels that are continuous in $\mathcal{I}$. Let the corresponding RKHS's be infinite dimensional. In the discretized representation, the kernel functions $\{K_i(s,t); s,t\in\mathcal{I}\}_{i=0}^{1}$ is approximated by $\mathbf{K}_i$, the corresponding $N\times N$ Gram matrices, whose elements are

$$\left(\mathbf{K}_i\right)_{mn} = K_i(t_n,t_m), \quad n,m = 1,2,\ldots N, \tag{95}$$

for $i=0,1$. Let $\{\nu_{ij}=\}_{j=1}^{N}$ be the (positive) eigenvalues of matrix $\mathbf{K}_i$. Theorem 3.4 of Baker (1977) can be used to show that, in the limit of dense monitoring,

$$\lim_{N\to\infty}\frac{\nu_j}{\Delta T} = \lambda_j, \quad j=1,2,\ldots,N \tag{96}$$

where $\{\lambda_{i1}\geq\lambda_{i2}\geq\ldots\geq\lambda_{iN}>0\}$ are the largest $N$ eigenvalues of $\mathcal{K}_i$, the covariance operator associated to the kernel $K_i$.

Therefore, the spectrum of the Gram matrix $\mathbf{K}_i$ converges to the spectrum of the covariance operator $\mathcal{K}_i$. In particular, the ratio of the determinants of the Gram matrix

$$\begin{aligned}\lim_{N\to\infty}\frac{|\mathbf{K}_1|}{|\mathbf{K}_0|} &= \lim_{N\to\infty}\prod_{j=1}^{N}\frac{\nu_{1j}}{\nu_{0j}}\\ &= \lim_{N\to\infty}\prod_{j=1}^{N}\frac{\lambda_{1j}}{\lambda_{0j}} \equiv \frac{|\mathcal{K}_1|}{|\mathcal{K}_0|},\end{aligned} \tag{97}$$

can be used to define the ratio $\frac{|\mathcal{K}_1|}{|\mathcal{K}_0|}$ when the corresponding Gaussian processes are equivalent ($\mathbb{P}_0\sim\mathbb{P}_1$), in which case the limit exists (is finite) and is different from zero.

## B Setup for the experiment with financial data

The setup of the experiment is as follows: Let $\{S_i(t_0),\ S_i(t_1),\ldots,S_i(t_L)\}$ be the time series of asset market prices for stock $i$ monitored at the equally-spaced instants

$t_n = t_0 + n\Delta T;\ n=0,1,\ldots,L,$

where $L = M(N_B + 1) - 1$. In the data analyzed $\Delta T$ is one day. Therefore, the quantity $S_i(t_n)$ is the closing price of the corresponding stock on the $n$th day of the period considered.

The time series is broken up into $M$ segments of length $N_B + 1$, with $N_B = 2^B$ for some integer $B$

$$\left\{S_i(t_0^{[m]}), S_i(t_1^{[m]}), \ldots, S_i(t_{N_B}^{[m]})\right\}_{m=1}^{M},$$

where $t_n^{[m]} = t_{n+(m-1)N_B}$, with $n = 0, 1, \ldots, N_B$, and $m = 1, 2, \ldots, M$. These $M$ time series of $N_B + 1$ prices are then transformed into the corresponding time series of log-returns

$$\left\{X_i(t_0^{[m]}), X_i(t_1^{[m]}), \ldots, X_i(t_{N_B}^{[m]})\right\}_{m=1}^{M}, \tag{98}$$

where

$$X_i(t_n^{[m]}) = \log \frac{S_i(t_n^{[m]})}{S_i(t_0^{[m]})}, \quad n = 0, 1, \ldots N_B.$$

The goal is to discriminate between different stocks on the basis of the corresponding time series of log-returns. In particular, we will analyze how the accuracy of the predictions depends on the monitoring frequency. For this reason, discrimination is made on the basis of $N_b + 1$ subsampled values within each segment

$$\left\{X_i(t_0^{[m]}), X_i(t_{n_b}^{[m]}), X_i(t_{2n_b}^{[m]}), \ldots, X_i(t_{N_b n_b}^{[m]})\right\},$$

where $N_b = 2^b$, and $n_b = 2^{B-b}$ with $b = 0, 1, \ldots, B$. As an illustration, for $b = 0$, only two inputs in each time series are used for discrimination

$$\left\{X_i(t_0^{[m]}), X_i(t_{N_B}^{[m]})\right\}.$$

For $b = B$ ($n_B = 1$) the complete time series given by Eq. (98) is used as input to the different classifiers. The higher monitoring the frequency is, the closer the problem is to a functional paradigm.

## References


Baíllo A, Cuevas A, Cuesta-Albertos JA (2011) Supervised Classification for a Family of Gaussian Functional Models. Scandinavian Journal of Statistics 38(3):480–498

Baker CTH (1977) The Numerical Treatment of Integral Equations. Clarendon, Oxford, U.K.

Berlinet A, Thomas-Agnan C (2004) Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer, Boston, MA, USA

Berrendero JR, Cárcamo J (2019) Linear components of quadratic classifiers. Advances in Data Analysis and Classification 13(2):347–377

Berrendero JR, Bueno-Larraz B, Cuevas A (2018a) On Mahalanobis distance in functional settings. arXiv:1803.06550

Berrendero JR, Cuevas A, Torrecilla JL (2018b) On the use of reproducing kernel Hilbert spaces in functional classification. Journal of the American Statistical Association 113(523):1210–1218

Bollerslev T, Chou R, Kroner KF (1992) Arch modeling in finance: A review of the theory and empirical evidence. Journal of Econometrics 52(1-2):5–59

Cont R (2001) Empirical properties of asset returns: stylized facts and statistical issues. Quantitative Finance 1(2):223–236

Cucker F, Smale S (2002) On the mathematical foundations of learning. Bulletin of the American Mathematical Society 39:1–49

Cucker F, Zhou DX (2007) Learning Theory: An Approximation Theory Viewpoint (Cambridge Monographs on Applied & Computational Mathematics). Cambridge University Press, New York, NY, USA

Cuesta-Albertos JA, Dutta S (2016) On perfect classification for Gaussian processes. arXiv:1602.04941

Cuevas A (2014) A partial overview of the theory of statistics with functional data. Journal of Statistical Planning and Inference 147:1 – 23

Dai X, Müller HG, Yao F (2017) Optimal Bayes classifiers for functional data and density ratios. Biometrika 104(3):545–560

Delaigle A, Hall P (2010) Defining probability density for a distribution of random functions. Ann Statist 38(2):1171–1193

Delaigle A, Hall P (2012) Achieving near perfect classification for functional data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74(2):267–286

Delaigle A, Hall P (2013) Classification using censored functional data. Journal of the American Statistical Association 108(504):1269–1283

Epifanio I, Ventura-Campos N (2014) Hippocampal shape analysis in Alzheimer's disease using functional data analysis. Statistics in Medicine 33(5):867–880

Fama EF (1965) The Behavior of Stock-Market Prices. The Journal of Business 38(1):34–105

Feldman J (1958) Equivalence and perpendicularity of Gaussian processes. Pacific J Math 8(4):699–708

Ferraty F, Vieu P (2006) Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics, Springer-Verlag New York, Inc., Secaucus, NJ, USA

Galeano P, Joseph E, Lillo RE (2015) The Mahalanobis distance for functional data with applications to classification. Technometrics 57(2):281–291, DOI 10.1080/00401706.2014.902774, URL https://doi.org/10.1080/00401706.2014.902774, https://doi.org/10.1080/00401706.2014.902774

Hájek J (1958) A property of $J$-divergences of marginal probability distributions. Czechoslovak Mathematical Journal 08(3):460–463

Hastie T, Tibshirani R, Friedman JH (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics, Springer

Hubert M, Rousseeuw P, Segaert P (2017) Multivariate and functional classification using depth and distance. Adv Data Anal Classif 11(3):445–466

Kailath T (1966) Some results on singular detection. Information and Control 9(2):130 – 152

Kailath T (1971) RKHS approach to detection and estimation problems–I: Deterministic signals in Gaussian noise. IEEE Transactions on Information Theory 17(5):530–549

Kuelbs J (1970) Gaussian measures on a Banach space. Journal of Functional Analysis 5(3):354 – 367

Leng X, Müller HG (2006) Classification using functional data analysis for temporal gene expression data. Bioinformatics 22(1):68–76

Lukić MN, Beder JH (2001) Stochastic Processes with Sample Paths in Reproducing Kernel Hilbert Spaces. Transactions of the American Mathematical Society 353(10):3945–3969

Manton JH, Amblard PO (2015) A primer on reproducing kernel Hilbert spaces. Foundations and Trends® in Signal Processing 8(1–2):1–126

Marks S, Dunn OJ (1974) Discriminant functions when covariance matrices are unequal. Journal of the American

Statistical Association 69(346):555–559
Martin-Barragan B, Lillo R, Romo J (2014) Interpretable support vector machines for functional data. European Journal of Operational Research 232(1):146 – 155
Müller HG (2016) Peter hall, functional data analysis and random objects. The Annals of Statistics 44(5):1867–1887
Osborne MFM (1959) Brownian motion in the stock market. Operations Research 7(2):145–173
Parzen E (1959) Statistical inference on time series by Hilbert space methods. Tech. rep., Tech. report 23, Statistics Department, Stanford University
Parzen E (1961a) An Approach to Time Series Analysis. Ann Math Statist 32(4):951–989
Parzen E (1961b) Regression analysis of continuous parameter time series. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, University of California Press, Berkeley, Calif., pp 469–489
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830
Ramos-Carreño C, Suárez A, Torrecilla JL, Carbajo Berrocal M, Marcos Manchón P, Pérez Manso P, Hernando Bernabé A (2019) scikit-fda: functional data analysis in Python. DOI 10.5281/zenodo.3468127, URL `https://doi.org/10.5281/zenodo.3468127`
Ramsay JO, Silverman BW (2005) Functional Data Analysis, 2nd edn. Springer Series in Statistics, Springer
Rasmussen CE, Williams CKI (2005) Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press
Rincón M, Ruiz-Medina MD (2012) Wavelet-RKHS-based functional statistical classification. Adv Data Anal Classif 6(3):201–217
Rossi F, Villa N (2006) Support vector machine for functional data classification. Neurocomputing 69(7):730 – 742, new Issues in Neurocomputing: 13th European Symposium on Artificial Neural Networks
Sato H (1967) On the equivalence of Gaussian measures. J Math Soc Japan 19(2):159–172
Shepp LA (1966) Radon-Nikodym derivatives of Gaussian measures. Ann Math Statist 37(2):321–354
Song JJ, Deng W, Lee HJ, Kwon D (2008) Optimal classification for time-course gene expression data using functional data analysis. Computational Biology and Chemistry 32(6):426 – 432
Spence A (1975) On the convergence of the Nyström method for the integral equation eigenvalue problem. Numerische Mathematik 25(1):57–66
Varberg DE (1961) On equivalence of Gaussian measures. Pacific J Math 11(2):751–762
Wahl PW, Kronmal RA (1977) Discriminant functions when covariances are unequal and sample sizes are moderate. Biometrics 33(3):479–484
Wang JL, Chiou JM, Müller HG (2016) Functional data analysis. Annual Review of Statistics and Its Application 3(1):257–295
Zhu H, Brown PJ, Morris JS (2012) Robust classification of functional and quantitative image data using functional mixed models. Biometrics 68(4):1260–1268